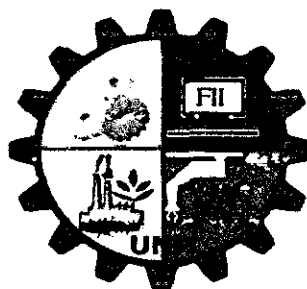


UNIVERSIDAD NACIONAL DE PIURA
FACULTAD DE INGENIERIA INDUSTRIAL
ESCUELA PROFESIONAL DE INGENIERIA INFORMÁTICA

PROYECTO DE TESIS



**“APLICACIÓN DE LA MINERÍA DE DATOS PARA ANALIZAR
LOS DIFERENTES TIPOS DE LESIONES REGISTRADOS EN LA
DIVISIÓN MÉDICO LEGAL – DISTRITO FISCAL DE PIURA.
PERÍODO 2005-2014”**

RESPONSABLES:

TESISTA

:
AURIESTELA DEL PILAR VICENTE LLACSAHUACHE

ASESOR

:
ING. NÉSTOR MANUEL CASTILLO BURGOS

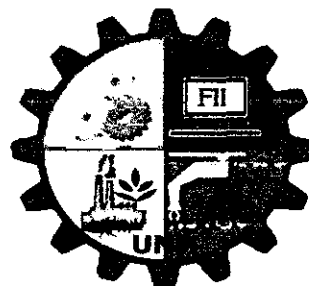
SEPTIEMBRE – 2015

7638

VIC



UNIVERSIDAD NACIONAL DE PIURA
FACULTAD DE INGENIERIA INDUSTRIAL
ESCUELA PROFESIONAL DE INGENIERIA INFORMÁTICA



**“APLICACIÓN DE LA MINERÍA DE DATOS PARA ANALIZAR
LOS DIFERENTES TIPOS DE LESIONES REGISTRADOS EN LA
DIVISIÓN MÉDICO LEGAL – DISTRITO FISCAL DE PIURA.**

PERÍODO 2005-2014”

**TESIS PARA OPTAR EL TÍTULO DE
INGENIERO INFORMÁTICO**

.....
AURIELSTELA DEL PILAR VICENTE LLACSAHUACHE

TESISTA

PIURA – PERU

2015

**“APLICACIÓN DE LA MINERÍA DE DATOS PARA ANALIZAR
LOS DIFERENTES TIPOS DE LESIONES REGISTRADOS EN LA
DIVISIÓN MÉDICO LEGAL – DISTRITO FISCAL DE PIURA.
PERÍODO 2005-2014”**

LINEA DE INVESTIGACIÓN

TESIS

**PRESENTADA A LA FACULTAD DE INGENIERIA INDUSTRIAL
DE LA UNIVERSIDAD NACIONAL DE PIURA
COMO REQUISITO PARA OPTAR EL TITULO DE:**

INGENIERO INFORMÁTICO

PROFESOR ASESOR

.....

ING. NÉSTOR MANUEL CASTILLO BURGOS

APROBADA

JURADO

PRESIDENTE

.....

FRANCISCO JAVIER CRUZ VILCHEZ

VOCAL

.....

CARMEN LUCILA INFANTE SAAVEDRA

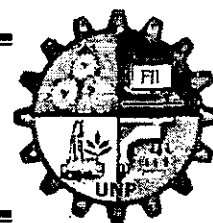
SECRETARIO

.....

JORGE LUIS SANDOVAL RIVERA



UNIVERSIDAD NACIONAL DE PIURA
FACULTAD DE INGENIERÍA INDUSTRIAL
DECANATO



ACTA DE SUSTENTACIÓN DE TESIS

Los Miembros del Jurado Calificador Ad-Hoc de la Tesis denominada: «**APLICACIÓN DE LA MINERÍA DE DATOS PARA ANALIZAR LOS DIFERENTES TIPOS DE LESIONES REGISTRADOS EN LA DIVISIÓN MÉDICO LEGAL – DISTRITO FISCAL DE PIURA, PERÍODO 2005 - 2014**», presentada por la señorita **AURIELA DEL PILAR VICENTE LLACSAHUACHE**, Bachiller de la Escuela Profesional en Ingeniería Informática; asesorada por el **Ing. Néstor Manuel Castillo Burgos** y co asesorada por el **Ing. Persi Williansh Cabrera Antón, MBA.**; reunidos para la sustentación de ésta y luego de escuchar su exposición y las respuestas a las preguntas formuladas, la declaran:



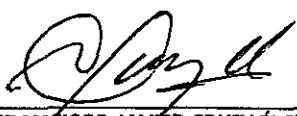
Con el Calificativo:

APROBADO


MUY BUENO

En consecuencia la sustentante se encuentra apta para recibir el título profesional de **INGENIERO INFORMÁTICO**, conforme a Ley.

Piura, 26 de setiembre del 2015


Ing. FRANCISCO JAVIER CRUZ VILCHEZ, MSc.
PRESIDENTE – JURADO CALIFICADOR


Ing. CARMEN LUCILA INFANTE SAAVEDRA, MSc.
VOCAL – JURADO CALIFICADOR


Ing. JORGE LUIS SANDOVAL RIVERA, MSc.
SECRETARIO – JURADO CALIFICADOR

DEDICATORIA

A Dios por acompañarme siempre.
Con todo amor a mis padres Elmer y
Primitiva por darme fuerzas para
seguir adelante.

A mis hermanos Fredy e Ibán
quienes me han regalado muchas
alegrías.

AGRADECIMIENTO

Debo agradecer a la Universidad Nacional de Piura que me permitió desarrollar mis estudios de Ingeniería Informática en tan prestigiosa institución.

RECONOCIMIENTO

A mi asesor Néstor y Co Asesor Persi, por brindarme la oportunidad de asesorarme en mi Proyecto de Investigación y optar el Título de Ingeniero Informático.

INDICE

INTRODUCCIÓN	10
RESUMEN	11
CAPITULO I: ASPECTOS GENERALES	12
1.1. Planteamiento del Problema	12
1.1.1. Enunciación del Problema	12
1.1.2. Formulación del Problema	14
1.2. Objetivos	14
1.2.1. Objetivo General	14
1.2.2. Objetivos Específicos	14
1.3. Justificación e importancia de la investigación	15
CAPÍTULO II: MARCO TEÓRICO	16
2.1. Antecedentes De La Investigación	16
2.2. Bases Teóricas	19
2.2.1. Definición de Lesiones	19
2.2.2. Tipos de Lesiones	19
2.2.3. Las lesiones en la División Médico Legal	19
2.2.4 Almacén De Datos (Data Warehouse), Objetivos Y Características	20
2.2.5. Arquitectura de un Almacén de Datos	24
2.2.6. Definición de Minería de Datos	29
2.2.7. Alcance de la Minería de Datos	30
2.2.8. Modelos de Minería de Datos	31
2.2.9. Arquitectura Del Modelo De Minería De Datos	31
2.2.10. Fases de la Minería de Datos	33
2.2.11. La Importancia de la Decisión	36
2.2.12. Sistemas de Apoyo para la Toma de Decisiones	38
2.2.13. Herramientas para la Toma de Decisiones	39
Capítulo III: METODOLOGÍA DE LA INVESTIGACIÓN.	41
3.1 Tipo, Nivel, Diseño y Método de Investigación	41
3.2 Método de Investigación	42

3.3	Cobertura del Estudio (Población Y Muestra)	42
3.4	Diseño de Técnicas e Instrumentos de Recolección de Información	42
3.5	Técnicas de Procesamiento, Análisis e Interpretación de Datos y/o Resultados	43
Capítulo IV: RESULTADOS Y DISCUSIÓN		44
4.1.	Selección de Datos para Elaborar Datawarehouse	44
4.2.	Diseño de las Estructuras del Datawarehouse	44
4.3.	Herramientas de Extracción, Transformación y Carga	47
4.3.1.	Paquete para extraer datos iniciales de la Fuente de Datos	47
4.3.2.	Paquete para crear la estructura Tipo de Lesiones, Ubicación Geográfica y la estructura Tiempo	50
4.4.	Modelo del Datawarehouse	54
4.5.	Modelos de Minería de Datos	55
4.5.1.	Algoritmo de Bayes Naive De Microsoft	55
4.5.2.	Algoritmo de Serie Temporal	66
4.5.3.	Algoritmo de Arboles de Decisión	76
CONCLUSIONES		88
RECOMENDACIONES		89
BIBLIOGRAFIA		90

INTRODUCCIÓN

Actualmente la sociedad peruana se encuentra muy convulsionada, siendo la ciudad de Piura una de las que mayor índice de violencia presenta, como producto de ésta violencia las personas involucradas en hechos de violencia provocan lesiones en otras personas o en algunas ocasiones sobre ellas mismas, hechos que son registrados diariamente en los sistemas de información de la División Médico Legal II de Piura y las Divisiones Médicos Legales I de Paita, Huancabamba, Chulucanas, Sechura y Catacaos. En la base de datos del sistema de información encontramos los diferentes tipos de lesiones por suceso de tránsito, lesiones por arma de fuego, lesiones por violencia familiar, lesiones múltiples, lesiones por arma blanca, lesiones por mordedura canina, otras lesiones.

El presente trabajo pretende a base de estos registros históricos ayudar a las autoridades encargadas de estos hechos violentos en el manejo de información relacionada con su trabajo con la finalidad de realizar análisis de la misma desde diversos puntos de vista y determinar acciones que crean convenientes para mejorar su gestión. La solución consiste en la obtención de modelos de Minería de Datos utilizando los algoritmos implementados en Microsoft Analysis Service, teniendo como base la implementación previa de un Datawarehouse alimentado por los datos almacenados en la base de datos de la División Médico Legal del Distrito Fiscal de Piura, para lo cual se utilizará las herramientas de extracción, transformación y carga del integrador de servicios de Microsoft Sql Server, de modo que los modelos obtenidos permitan el análisis de los registros de los diferentes tipos de lesiones, obtener patrones de comportamiento y hacer predicciones, de ésta forma dar la facilidad a las autoridades para que interpreten mejor dicha información; a través de una serie de elementos gráficos, flexibles en un entorno amigable que permita acceder de manera eficiente a la información de mejor calidad, más confiable para una buena toma de decisiones.

RESUMEN

El proyecto consiste en el desarrollo de modelos de Minería de datos para la toma de decisiones por parte de las autoridades sobre los hechos de violencia que causen lesiones físicas en las personas y se presenten en la región, el propósito de la implementación de dichos modelos es crear conocimiento que ayude a innovar los existentes. Así como mejorar la calidad de los servicios que presta la entidad,

En el capítulo I se alcanza las pautas generales de la investigación, como el planteamiento del problema, los objetivos alcanzados en la investigación, su justificación e importancia. En el capítulo II se presenta el marco teórico como los antecedentes de la investigación y las bases teóricas de la investigación. En el capítulo III se presenta la metodología de la investigación, donde se definió el tipo de investigación, el nivel de investigación, el método de investigación y el diseño de la investigación. Finalmente en el capítulo IV se presenta el resultado de la investigación que son los modelos de minería de datos obtenidos utilizando los algoritmos de Bayes Naive, Series de Tiempo y Árboles de Decisión, del mismo modo se discuten sus resultados con la interpretación de los resultados de los modelos.

CAPÍTULO I: ASPECTOS GENERALES

1.1. PLANTEAMIENTO DEL PROBLEMA

1.1.1. ENUNCIACIÓN DEL PROBLEMA

Actualmente nuestro país vive una convulsión social y Piura no es la excepción. Nuestra sociedad presenta diferentes problemas, estamos viviendo en una sociedad violenta, para darnos cuenta de ello es cuestión de salir a la calle, ser empujado o en el peor de los casos, incluso insultado por quien apenas accidentalmente rozamos, o también por las noches con el noticiero que narra hechos de violencia doméstica día tras día.

Además, la Organización Mundial de la Salud (OMS) considera un problema de salud pública los casos de mordedura canina, donde Piura es una de los departamentos con más casos presentados, estos casos de mordedura canina en el mundo son la causa de decenas de millones de lesiones cada año.

Otra situación problemática en nuestra sociedad son los accidentes de tránsito, un indicador de ello se ve claramente reflejado al afirmar que cada año, en el mundo mueren 1,2 millones de personas por causas de choques en las vías públicas y hasta 50 millones resultan heridas y más de 3000 personas mueren diariamente por lesiones resultantes del tránsito, esto es afirmado por el Consejo Regional de Seguridad Vial (CORESEVI) (Organización Mundial de la Salud, 2015, p.10) y según el Organismo Peruano de Consumidores y Usuarios (OPECU) en su cuarto informe vial informó que del lunes 25 de agosto al miércoles 24 de setiembre 2014, se han registrado 220 fallecidos en accidentes de tránsito. (Diario La Republica.pe, 2014)

Estos problemas sociales muchas veces es dada por trastornos de la personalidad, depresión, lo que causa en las personas irritabilidad, temperamento violento, falta de dominio, además de su imprudencia, falta de concientización e irresponsabilidad de cada persona involucrada en estos hechos, por otra parte también están incluidas las autoridades encargadas que no brindan la

atención y cuidado que la sociedad necesita, pues en el caso de los accidentes de tránsito algunas veces son causados por falta de mantenimiento de carreteras, señalización, falta de semáforos entre otros.

Las entidades encargadas como: La Fiscalía, la Policía, MINSA, el Ministerio del Interior, CORESEVI (en adelante Autoridades), en su intención por tratar de prever y minimizar estos hechos que en consecuencia producen en las personas afectadas leves o graves lesiones que ponen en peligro la vida de estos, solo se limitan a registrar los hechos mencionados según les corresponda, frente a ello deberían promover acciones de concientización, prevención pero la realidad problemática social se mantiene lo que muestra que no se está tomando acciones eficientes para contrarrestar esta convulsión social.

Esto significa que, en caso no mejoren las estrategias para reducir el alto índice de los problemas sociales mencionados seguiremos teniendo una sociedad mucho más conflictiva con un mayor registro de muertes, accidentes, lesiones.

Por ello, en el tema de seguridad ciudadana, con el desarrollo de este proyecto se pretende combatir frontalmente estos problemas sociales a través de dos aspectos: técnico y social. En el primero, hacer uso de las Tecnologías de Información; a través de la aplicación de Minería de Datos sobre los Registros Médico Legal (RML) históricos de los diferentes tipos de lesiones, estos registros serán tomados de la División Médico Legal (DML) del Distrito Fiscal de Piura con la intención de analizar los datos involucrados en cada RML de lesiones.

En el aspecto social, el presente proyecto permitirá realizar un trabajo coordinado entre la población y las autoridades, basado en la información analizada, las proyecciones brindadas después de aplicarse la Minería de Datos, lo que permitirá que las Entidades cambien y/o replanteen las estrategias para prever y/o minimizar los diferentes tipos de lesiones originados por hechos de violencia, mordeduras caninas y accidentes de tránsito.

1.1.2. FORMULACIÓN DEL PROBLEMA

¿De qué manera la implementación de la Tecnología de Minería de Datos ayudaría a analizar los registros de atención por diferentes tipos de lesiones en la División Médico Legal del Distrito Fiscal de Piura?

1.2. OBJETIVOS

1.2.1. OBJETIVO GENERAL

Implementar la Minería de Datos sobre los datos registrados en la División Médico Legal del Distrito Fiscal de Piura para analizar los diferentes tipos de lesiones

1.2.2. OBJETIVOS ESPECÍFICOS

1.2.2.1. Utilizar las herramientas de integración de datos para el proceso de extracción, transformación y carga de los datos.

1.2.2.2. Crear un almacén de datos (DATAWAREHOUSE), para realizar el desarrollo de modelos a través de la minería de datos.

1.2.2.3. Implementar el algoritmo de Bayes Naive para encontrar patrones de comportamiento en los diferentes tipos de lesiones

1.2.2.4. Implementar el algoritmo de series de tiempo para pronosticar en el tiempo los diferentes tipos de lesiones.

1.3. JUSTIFICACIÓN E IMPORTANCIA DE LA INVESTIGACIÓN

Debido a la convulsión social en la que Piura se encuentra sumergida, muestra de ello es el gran volumen histórico de Registros Médicos Legales de diferentes tipos de lesiones en la División Médico Legal del Distrito Fiscal de Piura, causados por armas de fuego, agresiones, accidentes de tránsito, mordeduras caninas entre otras.

El presente proyecto contribuye con la creación de conocimiento útil y pretende ayudar a las Autoridades encargadas como: La Policía, el Ministerio del Interior, La Fiscalía, Ministerio de Transporte, el MINSA entre otras en la planificación de sus actividades futuras frente a los hechos mencionados.

La aplicación de Minería de Datos sobre los datos de estos registros históricos ayudará en el análisis de estos hechos de violencia, a la toma de decisiones proactivas, análisis del comportamiento social en periodos anteriores, proyección del comportamiento social en un periodo próximo de dos años, concientización en la sociedad acerca de esta problemática social.

Además, permitirá determinar las causas más frecuentes de diferentes tipos de lesiones, ayudará a determinar cuáles son las zonas más vulnerables de sufrir algún tipo de lesión específica y por ende este análisis de información ayudará a un cambio de conciencia colectiva.

CAPÍTULO II: MARCO TEORICO

2.1. ANTECEDENTES DE LA INVESTIGACIÓN

- 2.1.1. Arturo Eguila Canales, Alex Alsey Parco Iquiapaza (Lima, 2007), en su tesis para optar el título profesional de Ingeniero de Sistemas denominada "Implementación de una herramienta de Inteligencia de Negocios para la administración de justicia sobre una metodología Ad-hoc", nos describe como El Sistema de Inteligencia de Negocios ofrece información confiable y oportuna en un ambiente gráfico, fácil de usar y con capacidad de análisis dinámico; esto ha permitido a los directivos dedicar más su tiempo a analizar información y tomar decisiones.¹
- 2.1.2. Buendia, M., 2014, en su trabajo de investigación "Algoritmos de Minería de Datos en la Recolección de Inteligencia" aplica algoritmos de MD en el ámbito delictivo con el fin de ayudar a prevenir hechos delictivos utilizando la metodología CRISP y algoritmos de agrupación como: K-means y también algoritmos predictivos como: árboles de decisión y Bayes Naives.
- 2.1.3. Britos, P., 2008, en su Tesis Doctoral "Procesos de Explotación de Información Basados en Sistemas Inteligentes". Esta tesis tiene como objetivo hacer aportaciones al cuerpo de conocimiento de los procesos de explotación de información y al cuerpo de conocimiento sobre nuevas aplicaciones de las tecnologías de sistemas inteligentes. Como conclusión se han obtenido futuras líneas de investigación, durante el desarrollo de la tesis se observó el uso indistinto de los términos de "minería de datos" y de "explotación de información" para referirse al mismo cuerpo de conocimiento. Sin embargo, plantear esta equivalencia es similar a plantear la equivalencia entre los sistemas informáticos y los sistemas de información. Los primeros describen la tecnología

¹ http://www.cybertesis.edu.pe/sdx/sisbib/fiche.xsp?base=documents&id=sisbib.2007.parco_ja-principal

que dan soporte a los segundos y esto es lo que los hace distintos. En este contexto surge como problema abierto de interés la necesidad de realizar un ordenamiento en el cuerpo de conocimiento en formación discriminando cuales son los procesos y las metodologías que pertenecen al campo de la explotación de información y cuáles son las tecnologías de minería de datos que dan soporte a dichos procesos y metodologías. Del mismo modo se concluye que en la literatura abundan los trabajos y resultados sobre la conveniencia de uso de determinados algoritmos de minería de datos frente a otros, sin embargo rara vez se plantea el proceso de explotación de información al cual estos algoritmos están asociados o la conveniencia del uso de uno algoritmo frente a otros en dicho proceso. En este contexto surge como problema abierto de interés la identificación de la correspondencia entre algoritmos de minería de datos y proceso de explotación de información.

- 2.1.4. Calderón, N., 2006, en su tesis “Minería de Datos una herramienta para la toma de decisiones” ayuda a evaluar el uso de Minería de Datos en la toma de decisiones.
- 2.1.5. Dhony Christian Espinoza Zevallos, Renan Rancisco Quispe Alvarez (Lima, 2006), en su tesis para optar el título profesional de Ingeniero de Sistemas denominada “Solución de inteligencia de negocios para empresas de servicios de asistencia aplicación práctica a la gerencia de asistencia del Touring y Automóvil Club del Perú”, El presente trabajo describe la implementación de una Solución de Inteligencia de negocios, como apoyo a la generación de conocimiento en razón a la información de la empresa, orientándose a empresas de Servicios de Asistencia desarrollando también una aplicación práctica en la Gerencia de Asistencia del Touring y Automóvil Club del Perú.

- 2.1.6. Escobar, J., 2007, en su Tesis Doctoral "Minería Web de Uso y Perfiles de Usuario: Aplicaciones con lógica Difusa". El objetivo del trabajo es centrarse en el análisis del comportamiento de los usuarios en la web, para poder obtener una representación de perfiles de usuarios. Como conclusión se podido obtener patrones de navegación y análisis demográfico (utilizando técnicas de clustering) de los usuarios que navegan por un determinado sitio web. A partir de estos perfiles de usuario se pueden llevar a cabo procesos de personalización para la mejora del sitio web
- 2.1.7. Espinoza, D., 2013, en su Tesis de Maestría "Análisis de la Empresa Agromar Industrial S.A. aplicando inteligencia de negocios", concluye que la finalidad de su trabajo es contribuir con el crecimiento del Grupo Agromar, a través de la creación de nuevo conocimiento producto de la aplicación de herramientas de inteligencia de negocios las cuales proporcionarán la información adecuada, en el momento preciso a los directivos y funcionarios para que las decisiones que éstos tomen sean las correctas. Para lograr lo anterior se debe unir a la estrategia de negocio, la estrategia tecnológica y de procesos, capaces de transformar los datos dispersos en conocimiento y de automatizar acciones comerciales para aumentar los ingresos de la empresa.
- 2.1.8. García, F., Cieza, J., Alvarado, B., 2005, en su artículo denominado "Revista Peruana de Medicina Experimental y Salud Publica" nos menciona a los accidentes de tránsito, agresiones y mordeduras caninas como un problema de Salud Pública.
- 2.1.9. Muñoz, F., 2012, en su artículo denominado "Mordedura canina" nos confirma que las mordeduras caninas son un grave problema de salud pública.
- 2.1.10. Romeu, P., 2010, en su trabajo de investigación "Minería de Datos aplicada al análisis del tratamiento informativo de la drogadicción"

da un enfoque en la aplicación de las técnicas de minería de datos haciendo uso de dos metodologías: KDD y CRISP.

2.2. BASES TEORICAS

2.2.1. DEFINICIÓN DE LESIONES

Daño o detrimento corporal causado por una herida, un golpe o una enfermedad. (Según la Real Academia Española - RAE).

En resumen de los conceptos dados en la Guía médico legal, se concluye que: Lesión es cualquier alteración física o psíquica que se causa a una persona, perjudicando su salud o integridad personal haciendo usos de agentes externos en un determinado tiempo y espacio. Además que toda lesión produce un daño, la tipificación jurídica de los hechos se realiza a través del Código Penal.

(Díaz, J. et al, 2014, pp.9-11)

2.2.2. TIPOS DE LESIONES

Las lesiones pueden ser producidas por diferentes agentes, para el estudio del proyecto se analizarán las lesiones producidas por los siguientes agentes causantes:

- Por agente contundente
- Por arma de fuego
- Por arma blanca
- Por accidentes de tránsito
- Por violencia familiar
- Por mordedura canina

(Díaz, J. et al, 2014, p.14)

2.2.3. LAS LESIONES EN LA DIVISIÓN MEDICO LEGAL

Los agentes causantes mencionados anteriormente muestran un problema social, cada uno de estos hechos son registrados en cada entidad competente. Las consecuencias de estos hechos producen lesiones en las partes afectadas es por ello

que se recurre a la División médico legal para determinar específicamente y la gravedad de la lesión producida.

Los Médicos Legistas realizan reconocimientos médicos legales a solicitud de las siguientes autoridades competentes tales como: Autoridades Judiciales, Fiscales, Policiales.

El Médico Legista debe consignar lo referido por el peritado, la situación o estado de hechos que motiva la atención, fecha, lugar, y naturaleza del mismo, el motivo del tipo de reconocimiento médico legal que se solicita (agresión física por violencia familiar, suceso de tránsito, suceso laboral o de trabajo, tortura, mordedura canina u otros). El Médico Legista, por su formación académica y experiencia, está en la capacidad de deslindar la parte subjetiva de este acápite, confrontando lo dicho por el peritado y los hallazgos producto del examen Clínico Forense.

(Díaz, J. et al, 2014, p.53)

2.2.4. ALMACÉN DE DATOS (DATA WAREHOUSE), OBJETIVOS Y CARACTERÍSTICAS

El Data Warehouse maneja un gran volumen de datos, debido a que consolida en su estructura la información recolectada durante años, proveniente de diversas fuentes y áreas, en un solo lugar centralizado. Es por esta razón que el depósito puede ser soportado y mantenido sobre diversos medios de almacenamiento.

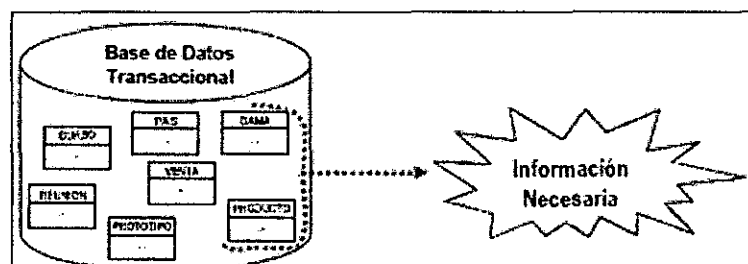
Organiza y almacena los datos que se necesitan para realizar consultas y procesos analíticos, con el propósito de responder a preguntas complejas y brindarles a los usuarios finales la posibilidad de que mediante una interface amigable, intuitiva y fácil de utilizar, puedan tomar decisiones sobre los datos sin tener que poseer demasiados conocimientos informáticos. El DW permite un acceso más directo, es decir, la información gira

en torno al negocio, y es por ello que también los usuarios pueden sentirse cómodos al explorar los datos y encontrar relaciones complejas entre los mismos

Debido a que W. H. Inmon, es reconocido mundialmente como el padre del DW, la explicación de las características más sobresalientes de esta herramienta se basó en su definición.

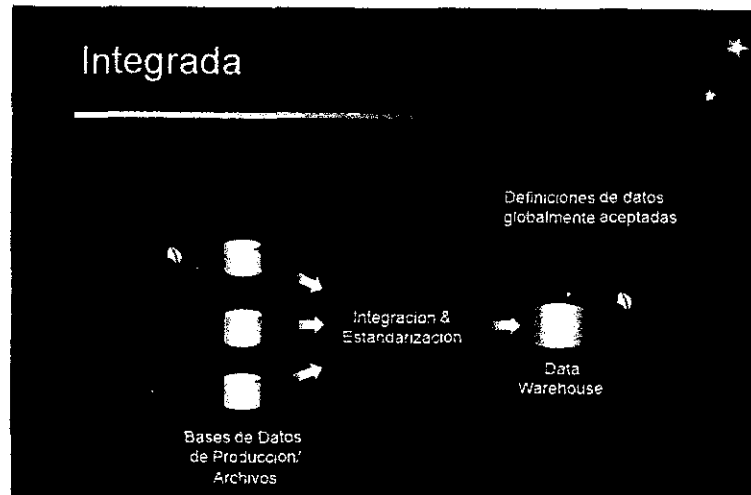
Orientada al negocio: La primera característica del DW, es que la información se clasifica en base a los aspectos que son de interés para la organización. Esta clasificación afecta el diseño y la implementación de los datos encontrados en el almacén de datos, debido a que la estructura del mismo difiere considerablemente a la de los clásicos procesos operacionales orientados a las aplicaciones.

Figura 2.1
Orientada hacia la información relevante de la organización



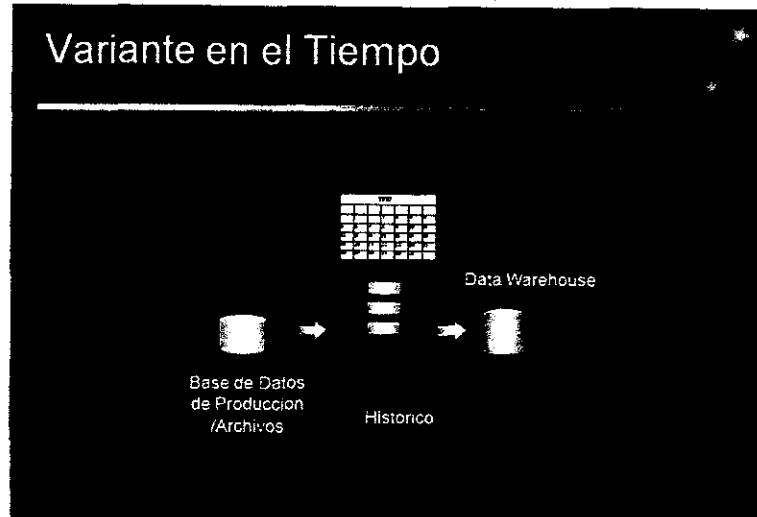
Integrada: La integración implica que todos los datos de diversas fuentes que son producidos por distintos departamentos, secciones y aplicaciones, tanto internos como externos, deben ser consolidados en una instancia antes de ser agregados al DW. La integración de datos, resuelve diferentes tipos de problemas relacionados con las convenciones de nombres, unidades de medidas, codificaciones, fuentes múltiples, etc.

Figura 2.2
Integración de Datos



Variante en el tiempo: Debido al gran volumen de información que se manejará en el DW, cuando se le realiza una consulta, los resultados deseados demorarán en originarse. Este espacio de tiempo que se produce desde la búsqueda de datos hasta su consecución es del todo normal en este ambiente y es, precisamente por ello, que la información que se encuentra dentro del depósito de datos se denomina de tiempo variable. El intervalo de tiempo y periodicidad de los datos debe definirse de acuerdo a la necesidad y requisitos de los usuarios. Es elemental aclarar, que el almacenamiento de datos históricos, es lo que permite al DW desarrollar pronósticos y análisis de tendencias y patrones, a partir de una base estadística de información.

Figura 2.3
Variante en el tiempo



No volátil: La información es útil para el análisis y la toma de decisiones solo cuando es estable. Los datos operacionales varían momento a momento, en cambio, los datos una vez que entran en el DW no cambian. La actualización, o sea, insertar, eliminar y modificar, se hace de forma muy habitual en el ambiente operacional sobre una base, registro por registro, en cambio en el depósito de datos la manipulación básica de los datos es mucho más simple, debido a que solo existen dos tipos de operaciones: la carga de datos y el acceso a los mismos. Por esta razón es que en el DW no se requieren mecanismos de control de la concurrencia y recuperación. (Bernabeu, 2009, pp.10-13)

Figura 2.4
No Volatilidad en los Datos

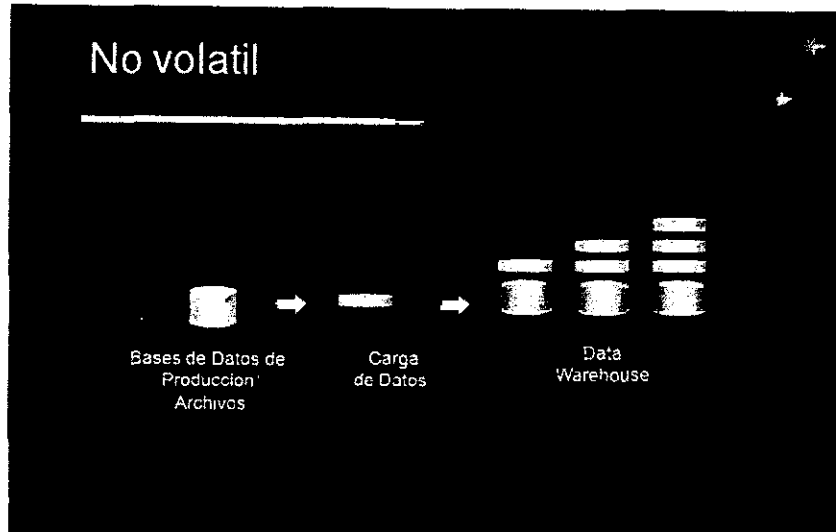
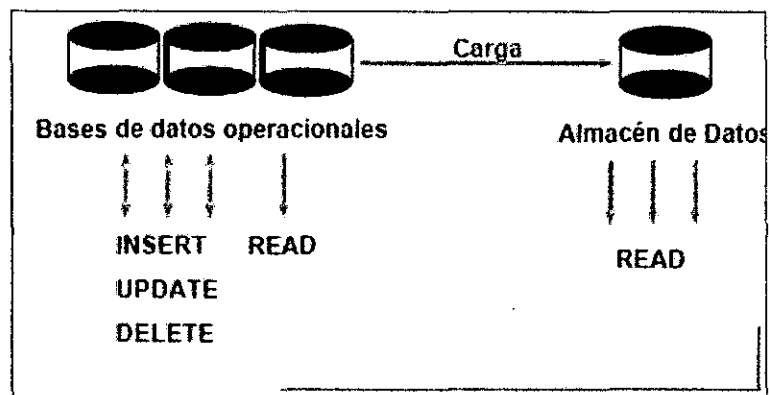


Figura 2.5
Operaciones del Data Warehouse (Carga y Lectura)



2.2.5. ARQUITECTURA DE UN ALMACÉN DE DATOS

A continuación se detallará cada uno de los componentes de la arquitectura del Data Warehousing.

OLTP (On Line Transaction Processing), representa toda aquella información transaccional que genera la empresa en su accionar diario, además, de las fuentes externas con las que puede llegar a disponer.

Entre los OLTP más habituales que pueden existir en cualquier organización se encuentran:

- ✓ Archivos de textos.
- ✓ Hipertextos.
- ✓ Hojas de cálculos.
- ✓ Informes semanales, mensuales, anuales, etc.
- ✓ Bases de datos transaccionales.

LOAD MANAGER: Para poder extraer los datos desde los OLTP, para luego manipularlos, integrarlos y transformarlos, para posteriormente cargar los resultados obtenidos en el DW, es necesario contar con algún sistema que se encargue de ello. Precisamente, los ETL (Extracción, Transformación y Carga) son los que cumplirán con tal fin.

Tal y como sus siglas lo indican, los ETL, extraen datos de las diversas fuentes que se requieran, los transforman para resolver posibles problemas de inconsistencias entre los mismos y finalmente, después de haberlos depurado se procede a su carga en el depósito de datos.

En síntesis, las funciones específicas de los ETL son tres: Extracción, Transformación y Carga.

A continuación, se detallará cada una de estas funciones, se expondrá cuál es el proceso que llevan a cabo los ETL y se enumerarán cuáles son sus principales tareas.

Extracción

Es aquí, en donde, basándose en las necesidades y requisitos del usuario, se exploran las diversas fuentes OLTP que se

tengan a disposición, y se extrae la información que se considere relevante al caso.

Si los datos operacionales residen en un SGBD Relacional, el proceso de extracción se puede reducir a, por ejemplo, consultas en SQL o rutinas programadas. En cambio, si se encuentran en un sistema no convencional o fuentes externas, ya sean textuales, hipertextuales, hojas de cálculos, etc, la obtención de los mismos puede ser un tanto más dificultoso, debido a que, por ejemplo, se tendrán que realizar cambios de formato y/o volcado de información a partir de alguna herramienta específica.

Transformación

Esta función es la encargada de convertir aquellos datos inconsistentes en un conjunto de datos compatibles y congruentes, para que puedan ser cargados en el DW. Estas acciones se llevan a cabo, debido a que pueden existir diferentes fuentes de información, y es vital conciliar un formato y forma única, definiendo estándares, para que todos los datos que ingresarán al DW estén integrados.

Además de lo antes mencionado, esta función se encarga de realizar el proceso de Limpieza de Datos (Data Cleansing). (Bernabeu, 2009, pp.20-22)

Carga

Este proceso es el responsable de cargar la estructura de datos del DW con:

- ✓ Aquellos datos que han sido transformados y que residen en el almacenamiento intermedio.

- ✓ Aquellos datos de los OLTP que tienen correspondencia directa con el depósito de datos.

Se debe tener en cuenta, que los datos antes de moverse al almacén de datos, deben ser analizados con el propósito de asegurar su calidad, ya que este es un factor clave, que no debe dejarse de lado. (Bernabeu, 2009, p.25)

Data Warehouse Manager

Gestiona el depósito de datos a través de tablas de hechos y tablas de dimensiones, y lo organiza en torno a una base de datos multidimensional³. Esto permite que se puedan crear cubos multidimensionales, Business Models u otras estructuras de datos. (Bernabeu, 2009, p. 27)

Query Manager

Este componente realiza las operaciones necesarias para soportar los procesos de gestión y ejecución de consultas relacionales, tales como Join y agregaciones, y de consultas propias del análisis de datos, como drill-up y drill-down.

Query Manager recibe las consultas del usuario, las aplica a la estructura de datos correspondiente (cubo multidimensional, Business Models, etc.) y devuelve los resultados obtenidos.

Cabe aclarar que una consulta a un DW, generalmente consiste en la obtención de indicadores a partir de los datos (hechos) de una tabla de hechos, restringidas por las propiedades o condiciones de los atributos que hayan sido creados.

Las operaciones que se pueden realizar sobre modelos multidimensionales y que son las que verdaderamente les permitirán a los usuarios explorar e investigar los datos en busca de respuestas, son:

- ✓ Drill-down.
- ✓ Drill-up.
- ✓ Drill-across.
- ✓ Roll-across.
- ✓ Pivot.
- ✓ Page.

(Bernabeu, 2009, p. 49)

Herramientas de Consulta y Análisis

Las herramientas de consulta y análisis son sistemas que permiten al usuario realizar la exploración de datos del DW. Básicamente constituyen el nexo entre el depósito de datos y los usuarios.

Utilizan la metadata de las estructuras de datos utilizadas (cubos multidimensionales, Business Models, etc.) para trasladar a través de consultas SQL los requerimientos del usuario, para luego, devolver el resultado obtenido.

A través de una amigable interfaz gráfica y una serie de simples pasos, el usuario genera consultas que son enviadas desde la herramienta de consulta y análisis al Query Manager, este a su vez realiza la extracción de información al DW Manager y devuelve los resultados obtenidos a la herramienta que se los solicitó. Luego, estos resultados son expuestos ante el usuario en formatos que le son familiares.

Una de las principales ventajas de utilizar estas herramientas, es que los usuarios no se tienen que preocupar por conocer cuáles son las características y funcionalidades de las estructuras de datos utilizadas, ni por saber emplear el lenguaje SQL, solo se deben enfocar en el análisis. (Bernabeu, 2009, pp. 61-62)

Usuarios

Los usuarios que posee el DW son aquellos que se encargan de tomar decisiones y de planificar las actividades del negocio, es por ello que se hace tanto énfasis en la integración, limpieza de datos, etc, para poder conseguir que la información posea toda la calidad posible.

Es a través de las herramientas de consulta y análisis, que los usuarios exploran los datos en busca de respuestas para poder tomar decisiones proactivas. (Bernabeu, 2009, p. 68)

2.2.6. DEFINICIÓN DE MINERÍA DE DATOS

La minería de datos se define dentro del proceso de descubrimiento de conocimiento en bases de datos como:

“La extracción de información interesante (no trivial, implícita, previamente desconocida y potencialmente útil), relaciones y/o patrones a partir de los datos en grandes bases de datos”. (Trujillo, J., Norberto, J., Pardillo, J., 2011, p. 82)

La minería de datos es el descubrimiento automático de patrones o modelos interesantes y no obvios escondidos en una base de datos, los cuales tienen un gran potencial para contribuir en los aspectos principales del negocio. La palabra interesante en los negocios se traduce en su aplicabilidad a las estrategias y tácticas de la empresa, como a sus objetivos.

Según Peacock la minería de datos también engloba la confirmación o prueba de relaciones reveladas por el proceso de descubrimiento, empleando para ello métodos estadísticos y bayesianos.

Como concepto más amplio, la minería de datos se identifica con el proceso de descubrir conocimientos en bases de datos (KDD KnowledgeDiscovery in Database).

(Barreiro, J., et al, 2003, p. 203)

La minería de datos es un paso en el proceso de KDD que consiste en la aplicación de análisis de datos y algoritmos de descubrimiento de que, en virtud de las limitaciones de eficiencia computacional aceptables, producen una enumeración particular de patrones (o modelos) en los datos. Se ocupa de los medios algorítmicos mediante el cual se extraen patrones de grandes cantidades de datos.

(AAAI, 1997, p.5)

Desde el punto de vista empresarial, se define como: La integración de un conjunto de áreas que tienen como propósito la identificación de un conocimiento obtenido a partir de las bases de datos que aporten un sesgo hacia la toma de decisión.

2.2.7. ALCANCE DE LA MINERÍA DE DATOS

La tecnología de Datamining puede generar nuevas oportunidades de negocios al proveer estas capacidades:

- Predicción automatizada de tendencias y comportamientos. Data Mining automatiza el proceso de encontrar información predecible en grandes bases de datos. Preguntas que tradicionalmente requerían un intenso análisis manual, ahora pueden ser contestadas directa y rápidamente desde los datos.

- Descubrimiento automatizado de modelos previamente desconocidos. Las herramientas de Datamining barren las bases de datos e identifican modelos previamente escondidos en un sólo paso. Otros problemas de descubrimiento de modelos incluye detectar transacciones fraudulentas de tarjetas de créditos e identificar datos anormales que pueden representar errores en la carga de datos.

(Vallejos, S., 2006, p.15-16)

2.2.8. MODELOS DE MINERÍA DE DATOS

Un modelo de minería de datos es un conjunto de datos, estadísticas y patrones que se pueden aplicar a los nuevos datos para generar predicciones y deducir relaciones. En función de su propósito general:

Modelos Descriptivos: describen el comportamiento de los datos de forma que sea interpretable por un usuario experto.

Modelos Predictivos: Además de describir los datos, se utilizan para predecir el valor de algún atributo desconocido.
(Perez, 2008, p.6)

2.2.9. ARQUITECTURA DEL MODELO DE MINERÍA DE DATOS

Un modelo de minería de datos recibe los datos de una estructura de minería de datos y, a continuación, los analiza utilizando un algoritmo de minería de datos. La estructura y el modelo de minería de datos son objetos independientes. La estructura de minería de datos almacena la información que define el origen de datos. Un modelo de minería de datos almacena la información derivada del procesamiento estadístico de los datos, como los patrones encontrados como resultado del análisis.

METADATOS: Los metadatos especifican el nombre del modelo y el servidor donde está almacenado, así como una definición del mismo, incluidas las columnas de la estructura de minería de datos que se usaron para generarlo, las definiciones de los filtros que se aplicaron al procesarlo y el algoritmo empleado para analizar los datos. Todas estas opciones, las columnas de datos y sus tipos de datos, filtros y algoritmo, influyen en gran medida en los resultados del análisis.

Los metadatos han de ser escalables y siempre están expuestos a cambios. Por ello, deben permitir la integración, registrar la historia y contenido del almacén de datos y deben reflejar los cambios que se realicen. Es fundamental para el proyecto la definición de una buena estrategia de metadatos que involucre la implantación, el control y la gestión de los metadatos en todas las fases del proyecto del almacén de datos. (Trujillo, J., Norberto, J., Pardillo, J., 2011, pp. 91-92)

PATRONES: El modelo también se ve afectado por los datos usados en el entrenamiento: incluso los modelos cuyo entrenamiento se ha realizado en la misma estructura de minería de datos pueden producir resultados distintos si se filtran los datos de manera diferente o se usan semillas distintas durante el análisis. Sin embargo, los datos reales no se almacenan en el modelo; solo lo hacen las estadísticas de resumen, mientras que los datos reales residen en la estructura de minería de datos. Si ha creado filtros en los datos al realizar el entrenamiento del modelo, las definiciones de filtro también se guardan con el objeto de modelo.

Es una expresión en un lenguaje que describe un subconjunto de datos o un modelo aplicable al subconjunto. Por lo tanto, un patrón también está designado a ajustar un modelo a los datos;

descubrir la estructura de datos; o, en general, hace cualquier descripción de alto nivel de un conjunto de datos.

(AAAI, 1997, p.5)

2.2.10. FASES DE LA MINERÍA DE DATOS

Según las bibliografías consultadas, se establece de manera genérica las siguientes fases para realizar un proyecto de Minería de Datos:

a) DETERMINACIÓN DE LOS OBJETIVOS

Este paso incluye analizar los requisitos empresariales, definir el ámbito del problema, definir las métricas por las que se evaluará el modelo y definir los objetivos concretos del proyecto de minería de datos.

b) ADQUISICIÓN Y EVALUACIÓN DE LOS DATOS

Se refiere a la selección, la limpieza (de manera que se eliminan valores incorrectos, no válidos según las necesidades y el algoritmo a usar), el enriquecimiento, la reducción y la transformación de las bases de datos, debido a que el formato de los datos contenidos en la fuente de datos nunca es el idóneo, y la mayoría de las veces no es posible ni siquiera utilizar ningún algoritmo de minería sobre los datos en bruto.

Consiste en consolidar y limpiar los datos identificados en el paso anterior. Los datos pueden estar dispersos en la empresa y almacenados en formatos distintos; también pueden contener incoherencias como entradas que faltan o incorrectas.

La limpieza de datos no solamente implica quitar los datos no válidos o interpolar valores que faltan, sino también buscar las correlaciones ocultas en los datos, identificar los

orígenes de datos que son más precisos y determinar qué columnas son las más adecuadas para el análisis.

c) EXPLORAR LOS DATOS O SELECCIÓN DE VARIABLES

Aún después de haber sido pre procesados los datos, en la mayoría de los casos se tiene una cantidad ingente de datos. La selección de características reduce el tamaño de los datos eligiendo las variables más influyentes en el problema, sin apenas sacrificar la calidad del modelo de conocimiento obtenido del proceso de minería.

Los métodos para la selección de características son básicamente dos:

- Aquellos basados en la elección de los mejores atributos del problema,
- Y aquellos que buscan variables independientes mediante test de sensibilidad, algoritmos de distancia o heurísticos.

Debe conocer los datos para tomar las decisiones adecuadas al crear los modelos de minería de datos.

Al explorar los datos para conocer el problema empresarial, puede decidir si el conjunto de datos contiene datos defectuosos y, a continuación, puede inventar una estrategia para corregir los problemas u obtener una descripción más profunda de los comportamientos que son típicos del negocio.

En conclusión, el objetivo de esta etapa es: Producir un conjunto de datos (data set) representativo, reproducible y confiable.

d) GENERAR MODELOS O ALGORITMOS DE EXTRACCIÓN DE CONOCIMIENTO

Desarrollar modelos descriptivos y/o predictivos, que representa patrones de comportamiento observados en los valores de las variables del problema o relaciones de asociación entre dichas variables.

También pueden usarse varias técnicas a la vez para generar distintos modelos, aunque generalmente cada técnica obliga a un pre procesamiento diferente de los datos.

e) EXPLORAR Y EVALUAR LOS MODELOS

Antes de implementar un modelo en un entorno de producción, aconsejan probar si funciona correctamente. Además, al generar un modelo, normalmente se crean varios con configuraciones diferentes y se prueban todos para ver cuál ofrece los resultados mejores para su problema y sus datos.

Existen herramientas que ayudan a separar los datos en conjuntos de datos de entrenamiento y pruebas, para que pueda evaluar con precisión el rendimiento de todos los modelos en los mismos datos. El conjunto de datos de entrenamiento se utiliza para generar el modelo y el conjunto de datos de prueba para comprobar la precisión del modelo mediante la creación de consultas de predicción.

También puede comprobar si los modelos crean predicciones correctamente mediante herramientas del diseñador como el gráfico de mejora respecto al modelo predictivo y la matriz de clasificación. Para comprobar si el modelo es específico de sus datos o se puede utilizar para realizar inferencias en la población general, se puede utilizar la técnica estadística denominada validación cruzada para crear automáticamente subconjuntos de los datos y probar el modelo con cada uno.

f) IMPLEMENTAR

Una vez obtenido el modelo, se debe proceder a su validación, comprobando que las conclusiones que arroja son válidas y suficientemente satisfactorias.

En el caso de haber obtenido varios modelos mediante el uso de distintas técnicas, se deben comparar los modelos en busca de aquel que se ajuste mejor al problema.

Una vez que los modelos de minería de datos se encuentran en el entorno de producción, se puede llevar a cabo diferentes tareas, dependiendo de las necesidades. Las siguientes son algunas de las tareas que se puede realizar:

- Usar los modelos para crear predicciones que luego se podrá usar para tomar decisiones comerciales.
 - Crear consultas de contenido para recuperar estadísticas, reglas o fórmulas del modelo.
 - Incrustar la funcionalidad de minería de datos directamente en una aplicación.
 - Crear un informe que permita a los usuarios realizar consultas directamente en un modelo de minería de datos existente.
 - Actualizar los modelos después de la revisión y análisis. Cualquier actualización requiere que se vuelve a procesar los modelos
 - Actualizar dinámicamente los modelos, cuando entren más datos en la organización, y realizar modificaciones constantes para mejorar la efectividad de la solución debería ser parte de la estrategia de implementación.
-
- (Pérez, 2008, pp.15-17)

2.2.11. LA IMPORTANCIA DE LA DECISION

A continuación se describen dos enfoques para decidir:

La Visión Determinista:

Es importante subrayar que el enfoque determinista ha tenido importancia en los principales estudios sobre la teoría de la decisión. Lucien Sfez, un experto en el tema, dice que la

decisión es un elemento ideológico que pretende dar movilidad a los sistemas políticos y administrativos. Su interpretación nos indica que la primera función que cumple la decisión es permitir que la persona actúe. Según este autor, si los que toman decisiones tuviesen presente en todo momento el peso de los determinismos y de las estructuras su "impulso hacia la acción se cortarí". La decisión por lo tanto, crea la ilusión de que son actores libres y que realmente están eligiendo autónomamente cursos de acción. La realidad, apunta Sfez, es que hay un aparato y una racionalidad que ya determinó el camino a seguir.

Otra de las funciones que, según Sfez, cumple la decisión en el mundo moderno es la de servir como válvula de escape de un universo hiperinstitucionalizado en el que todo tiene un procedimiento, una norma, un código o un sistema de coerción, y en el cual la mayoría de nuestros actos y nuestras voliciones tienen una forma algorítmica o predeterminada. es decir, tenemos la ilusión de ser libres pero en realidad no lo somos.

La otra cara de la moneda: El Enfoque Liberal

La contraparte de la visión determinista se basa en la idea de un hombre libre para decidir su circunstancia y construir su futuro. Esta corriente se basa en tres pilares cartesianos que son:

- a) Lógica lineal
- b) Racionalidad
- c) Libertad

La linealidad tiene que ver con un principio causa - efecto: si hago x por ejemplo, debo obtener y, de acuerdo con la experiencia y con el saber disponibles en un momento histórico determinado.

La racionalidad es el elemento que orienta la decisión, se habla de racionalidad religiosa, económica, política, administrativa, etcétera. Es decir, un conjunto de normas y de fines que se explican de manera autónoma y hacen operante la relación lineal causa - efecto.

El tercer elemento, la libertad del hombre para elegir, se funda en que toda persona posee una dotación natural de atributos, voluntades y capacidades que le permiten elegir entre las diferentes opciones de acuerdo con su libre albedrío.

(Curzio, 1998, pp 18-19)

2.2.12. SISTEMAS DE APOYO PARA LA TOMA DE DECISIONES

Una de las clasificaciones de los sistemas de información en las organizaciones son los llamados Sistemas de Apoyo para la Toma de Decisiones (SAD) los cuales se describen como:

Los SAD no son estructurados y formalizados puesto que generalmente se usan para tratamientos ad hoc y, por tanto, la flexibilidad y capacidad de adaptación son cruciales.

El énfasis se sitúa en la labor de soporte y no en la automatización de las decisiones. El ordenador debe apoyar al decisor facilitando el acceso a los datos y la prueba de soluciones alternativas pero no debe remplazar su criterio, es decir, no debe tratar de dar respuestas ni imponer una secuencia de análisis predefinida. Es el hombre quien en último término toma la decisión e, incluso en el caso de que el sistema de ayuda se asocie a un sistema experto, será consultado como lo sería un experto humano y el decisor podrá aceptar o rechazar las opciones propuestas.

El sistema puede almacenar y posteriormente reprocesar sus propios modelo de datos. El usuario interactúa con el sistema: realiza demandas, crea o modifica modelos para adaptados a las variaciones en su comprensión del problema y a sus necesidades, manipula datos y diseña el formato y contenido de los informes a obtener los cuales pueden incluir texto, informaciones estructuradas o gráficos.

(Sandoval, M. & González, E., 1995, pp. 140-141)

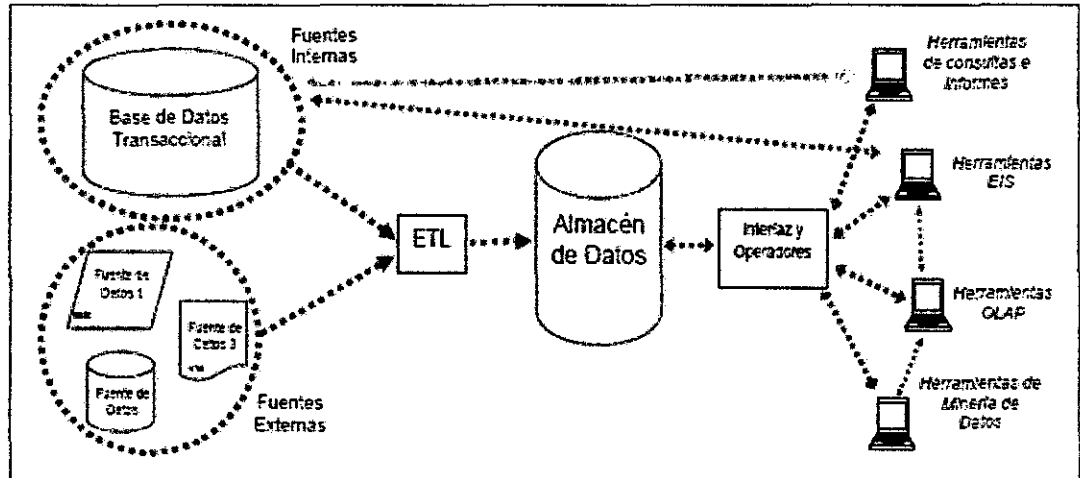
2.2.13. HERRAMIENTAS PARA LA TOMA DE DECISIONES

Han aparecido diferentes herramientas de negocio para la toma de Decisiones (DSS o DecisionSupportSystems) que coexisten: EIS OLAP, consultas e informes, y las propias herramientas de minería de datos.

Las herramientas OLAP proporcionan facilidades para manejar y transformar los datos, producen otros datos (más agregados, combinados) y son una gran ayuda para analizar los datos porque producen diferentes vistas de los mismos.

Las herramientas de Minería de Datos son muy variadas: permiten "extraer" patrones, modelos, descubrir relaciones, regularidades, tendencias, para describir y comprender mejor los datos y para predecir comportamientos futuros. Producen "reglas" o "patrones" ("conocimiento").

Figura 2.6
Interrelaciones entre herramientas de Sistemas Soporte Decisiones



Las herramientas anteriormente citadas suelen necesitar de la existencia previa de un almacén de datos (Data Warehouse). El almacén de datos es el sistema de información central en todo éste proceso.

(Pérez, C. & Santín, D., 2008, p. 2-3)

CAPÍTULO III: METODOLOGÍA DE LA INVESTIGACIÓN

3.1. TIPO, NIVEL, DISEÑO Y MÉTODO DE INVESTIGACIÓN

3.1.1. TIPO DE ORIENTACIÓN

Es un tipo de orientación mixta, es decir tiene de investigación básica porque se van a incrementar los conocimientos y es aplicada porque busca la aplicación o utilización de los conocimientos que se adquieren.

3.1.2. NIVEL DE INVESTIGACIÓN

Es un nivel de investigación descriptivo.

3.1.3. DISEÑO DE INVESTIGACIÓN

Corresponde al tipo de investigación Descriptiva y Longitudinal. Se considera investigación descriptiva aquella en que, como afirma Salkind, "se reseñan las características o rasgos de la situación o fenómeno objeto de estudio".

Para Cerda "tradicionalmente se define la palabra describir como el acto de representar, reproducir o figurar a personas, animales o cosas..."; y agrega: "Se deben describir aquellos aspectos más característicos distintivos y particulares de estas personas, situaciones o cosas, osea, aquellas propiedades que las hacen reconocibles a los ojos de los demás".²

Según Briones, en relación con el tiempo o número de veces en que se obtiene información del objeto de estudio, existen dos tipos de investigación: la seccional y la longitudinal. En la investigación longitudinal se obtienen datos de la misma población en distintos momentos durante un periodo determinado, con el fin de examinar sus variaciones en el tiempo³. Ejemplo, un estudio a evaluar la productividad de los trabajadores de una empresa y para ello se

² CERDA, Hugo y BRIONES Guillermo (Sin fecha). Manual Metodología de la Investigación. Venezuela, P. 116.

³ CERDA, Hugo y BRIONES Guillermo (Sin fecha). Manual Metodología de la Investigación. Venezuela, P. 152.

analiza la producción durante dos (2) años, recogiendo información tres veces por año.

3.2. MÉTODO DE INVESTIGACIÓN

Se empleará el método científico como método general y, como específicos al método analógico e inferencial.

3.3. COBERTURA DEL ESTUDIO (POBLACIÓN Y MUESTRA)

3.3.1. POBLACIÓN

La población del proyecto será el total de los registros de los diferentes tipos de lesiones registrados en el Distrito Fiscal de Piura; es decir, Piura, Catacaos, Sechura, Paita y Chulucanas.

3.3.2. MUESTRA

Para efectos del presente trabajo de investigación se trabajará sobre toda la población.

3.4. DISEÑO DE TÉCNICAS E INSTRUMENTOS DE RECOLECCIÓN DE INFORMACIÓN

Las técnicas de recolección de datos que se utilizará es el análisis documental mediante documentos escritos, documentos estadísticos, y la técnica utilizada en el análisis documental es: La técnica de lectura, la técnica de observación.

Los instrumentos que se utilizarán son: La lista de cotejo, porque me permite investigar a través del cuadro de doble entrada, que datos (fila horizontal) necesito de los documentos estadísticos u otros documentos escritos analizados (fila vertical), y, la entrevista, porque permite conocer las necesidades de información de los ejecutivos y de acuerdo con ello recolectar los datos que satisfagan esas necesidades.

3.5. TÉCNICAS DE PROCESAMIENTO, ANÁLISIS E INTERPRETACIÓN DE DATOS Y/O RESULTADOS.

Los datos cuantitativos serán procesados y analizados por medios computarizados, clasificados y sistematizados de acuerdo a las unidades de análisis correspondientes, respecto a sus variables a través del software de Analisys Service.

CAPÍTULO IV: RESULTADOS Y DISCUSIÓN

4.1. SELECCIÓN DE DATOS PARA ELABORAR DATAWAREHOUSE.

Los datos que se utilizan para el presente trabajo tiene como fuente de origen las estadísticas publicadas por el Ministerio Público, esto se da por la confidencialidad de la información que es propia de la institución y como consecuencia de ello no se tiene acceso a los datos almacenados en sus base de datos. Por lo tanto los datos a utilizar para aplicar las herramientas de minería de datos son los siguientes:

- Lesiones por proyectil arma de fuego
- Lesiones por suceso de tránsito
- Lesiones por mordedura canina
- Lesiones por violencia familiar
- Lesiones por arma blanca
- Lesiones por accidente de trabajo
- Lesiones por trata de personas
- Lesiones por presunto delito contra la libertad sexual
- Lesiones múltiples
- Lesiones a detenidos
- Otras lesiones

4.2. DISEÑO DE LAS ESTRUCTURAS DEL DATAWAREHOUSE

Para diseñar el Datawarehouse se ha tomado la data registrada en la base de datos de la División Médico Legal del Distrito Fiscal de Piura. La población del presente trabajo de investigación está constituida por los datos de los Registros Médicos legales de la División Médico Legal del Distrito Fiscal de Piura; que incluye: Piura, Catacaos, Sechura, Paita y Chulucanas.

Estructuras utilizadas en el datawarehouse

Estructura GÉNERO

Atributos	Contenido		
	Descripción	Formato	Valor por defecto
Idgenero	Llave Primaria	Nvarchar(1)	Ninguno
Sexo	Descripción del sexo de la persona	Nvarchar(30)	Ninguno

Estructura LESIONES

Atributos	Contenido		
	Descripción	Formato	Valor por defecto
Idtipolesion	Llave Foránea de la Estructura TipoLesiones	Int	Ninguno
Edad	Edad de la persona registrada	Int	Ninguno
Idgenero	Llave Foránea de la Estructura Genero	Nvarchar(1)	Ninguno
Clave	Llave Primaria	Int	Ninguno
Indice	Llave Foránea de la Estructura Tiempo	char(8)	Ninguno

Estructura TIEMPO

Atributos	Contenido		
	Descripción	Formato	Valor por defecto
Indice	Llave Primaria de la Dimensión Tiempo	char(8)	Ninguno
Año	El año obtenido de la Fecha	char(4)	Ninguno
Semestre	El semestre obtenido de la Fecha	char(10)	Ninguno
Trimestre	El trimestre obtenido de la Fecha	char(10)	Ninguno
Bimestre	El bimestre obtenido de la Fecha	char(10)	Ninguno
Mes	El mes obtenido de la Fecha	char(2)	Ninguno
Dia	El día obtenido de la Fecha	char(2)	Ninguno

Estructura TIPOLESIONES

Atributos	Contenido		
	Descripción	Formato	Valor por defecto
IdTipoLesion	Llave Primaria	Int	Ninguno
DescripcionLesion	Denominación del Tipo de Lesión	varchar(100)	Ninguno

Estructura UBIGEO

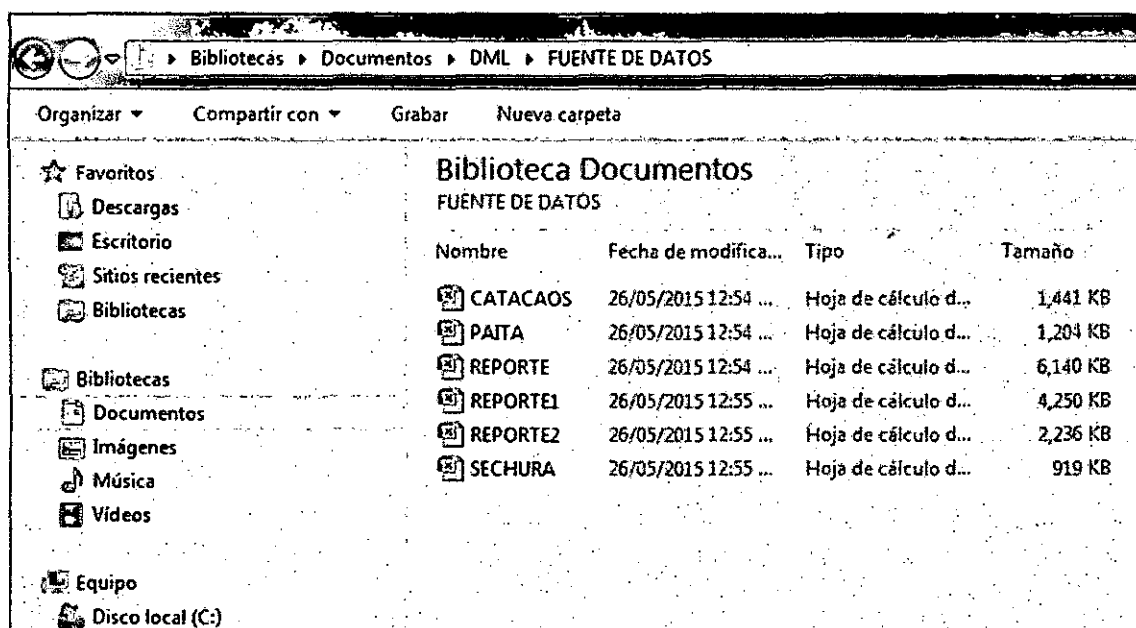
Atributos	Contenido		
	Descripción	Formato	Valor por defecto
IdDml	Llave Primaria	Int	Ninguno
DescripcionDml	Descripción de la División Médico Legal	Varchar (80)	Ninguno

4.3. HERRAMIENTAS DE EXTRACCIÓN, TRANSFORMACIÓN Y CARGA

4.3.1. Paquete para extraer datos iniciales de la Fuente de Datos (Lesiones.dtsx)

Paso 01: Identificadas y seleccionadas las fuentes de datos que van a ser analizadas (Enero 2006 - Abril 2015) y posteriormente explotadas con el objetivo de hacer predicciones y obtener patrones de comportamiento, estas fuentes de datos se encuentran en formato Excel versión 2007, las cuales se colocan en el servidor en la ruta C:\Users\auri\Documents\DML\FUENTE DE DATOS. Tal como se indica en la figura 4.1

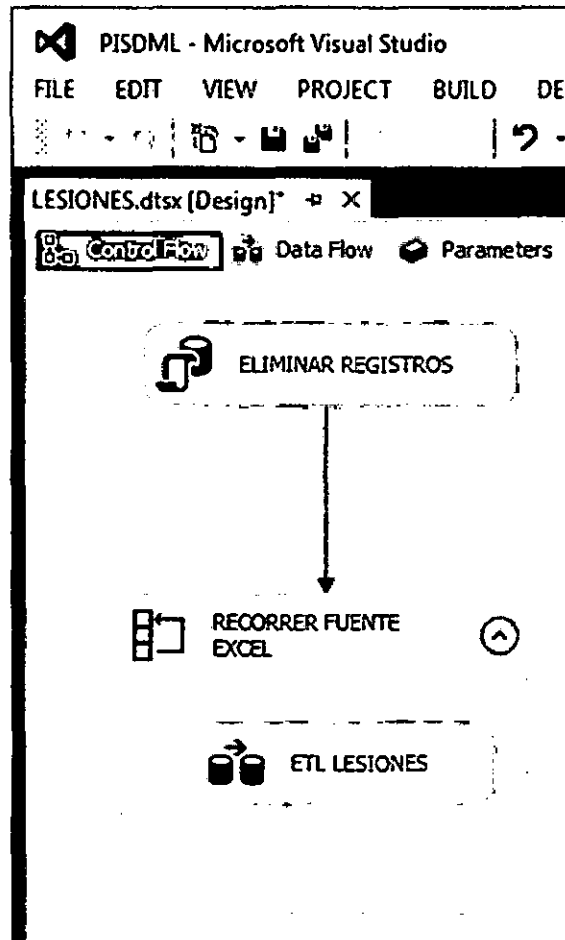
Fig. 4.1 Fuentes de Datos en Excel



Paso 02: Luego de hacer una exploración previa a los datos del origen en archivo Excel, se procede a elaborar un paquete utilizando la herramienta de extracción transformación y carga de Integration Service de Microsoft Sql Server 2014. En el paquete elaborado se hace uso de los controles: execute sql task (ejecutar tarea sql), foreach loop container (bucle

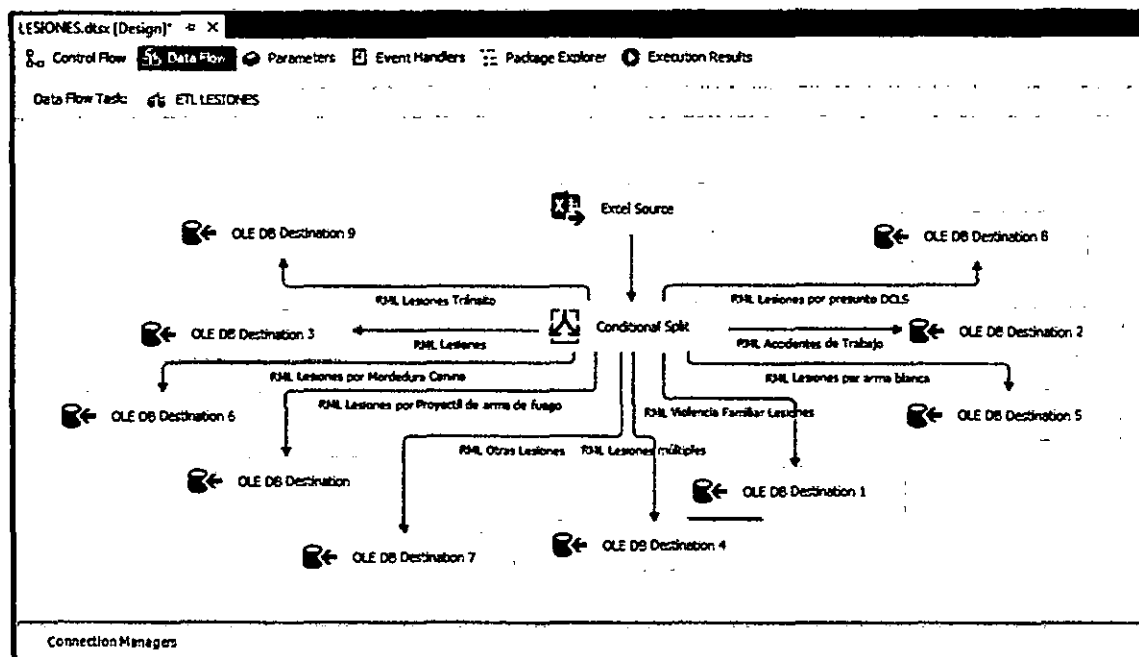
contenedor) y data flow task (tarea de flujo de datos), tal como se visualiza en la figura 4.2

Fig. 4.2 Controles del paquete



Continuando con la descripción del paquete elaborado se han utilizado controles de transformación de datos en el flujo de datos como conditional split y controles de origen de datos como Excel source y controles de destino de datos como ole db destination, tal como se visualiza en la figura 4.3

Fig. 4.3 Controles del flujo de datos



Para una mejor comprensión del control conditional split Figura 4.4 se detalla cada uno de los casos condicionales que se han implementado para filtrar los datos de las fuentes de datos de origen en Excel al destino que es el Datawarehouse implementado en sql server.

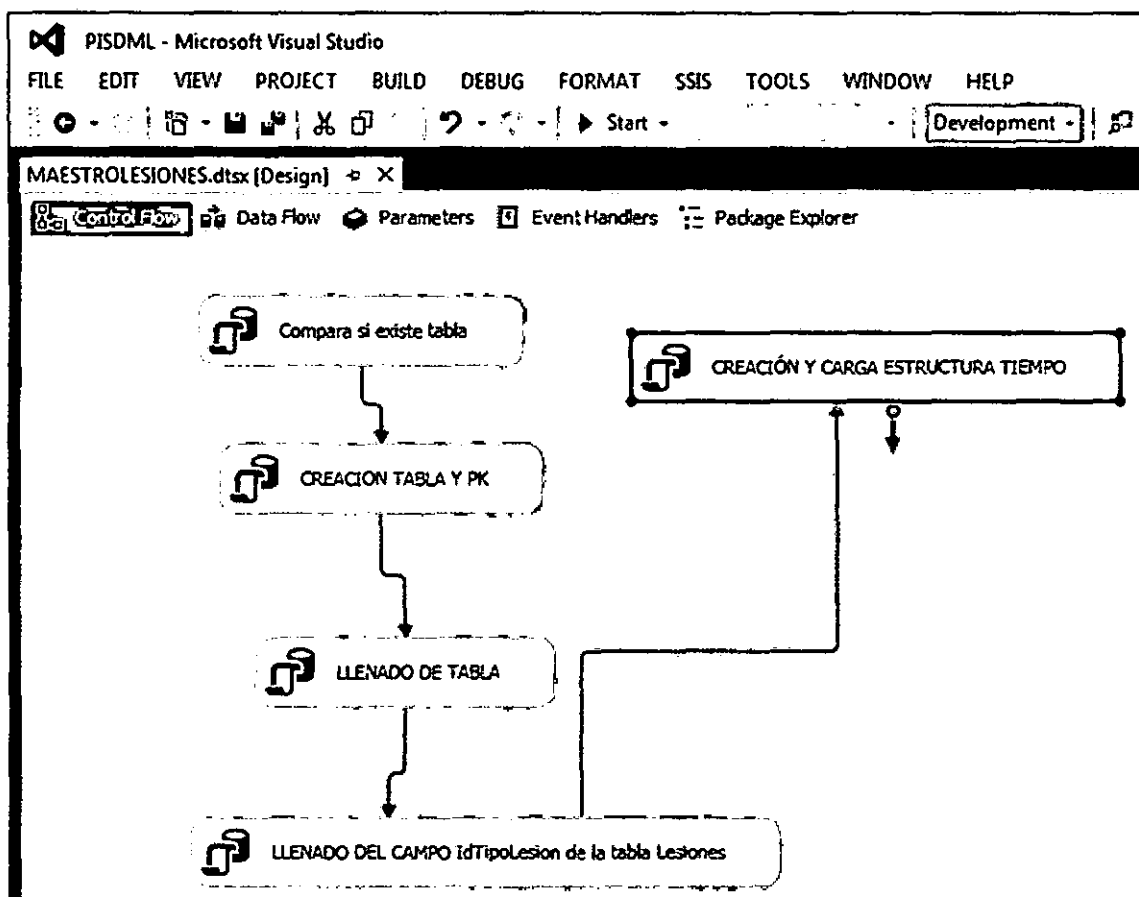
Fig. 4.4 Control conditional split

Conditional Split Transformation Editor		
Specify the conditions used to direct input rows to specific outputs. If an input row matches no condition, the row is directed to a default output.		
<div> <div>Variables and Parameters</div> <div>Columns</div> </div>		<div> <div>Mathematical Functions</div> <div>String Functions</div> <div>Date/Time Functions</div> <div>NULL Functions</div> <div>Type Casts</div> <div>Operators</div> </div>
Order	Output Name	Condition
1	RML Lesiones por Proyectoil de arma de fuego	de_serv == "RML Lesiones por Proyectoil de arma de fuego"
2	RML Violencia Familiar Lesiones	de_serv == "RML Violencia Familiar Lesiones"
3	RML Accidentes de Trabajo	de_serv == "RML Accidentes de Trabajo"
4	RML Lesiones por presunto DCLS	de_serv == "RML Lesiones por presunto DCLS"
5	RML Otras Lesiones	de_serv == "RML Otras Lesiones"
6	RML Lesiones Tránsito	de_serv == "RML Lesiones Tránsito"
7	RML Lesiones por Mordedura Canina	de_serv == "RML Lesiones por Mordedura Canina"
8	RML Lesiones por arma blanca	de_serv == "RML Lesiones por arma blanca"
9	RML Lesiones múltiples	de_serv == "RML Lesiones múltiples"
10	RML Lesiones	de_serv == "RML Lesiones"

4.3.2. Paquete para crear la estructura Tipo de Lesiones, Ubicación Geográfica y la estructura Tiempo

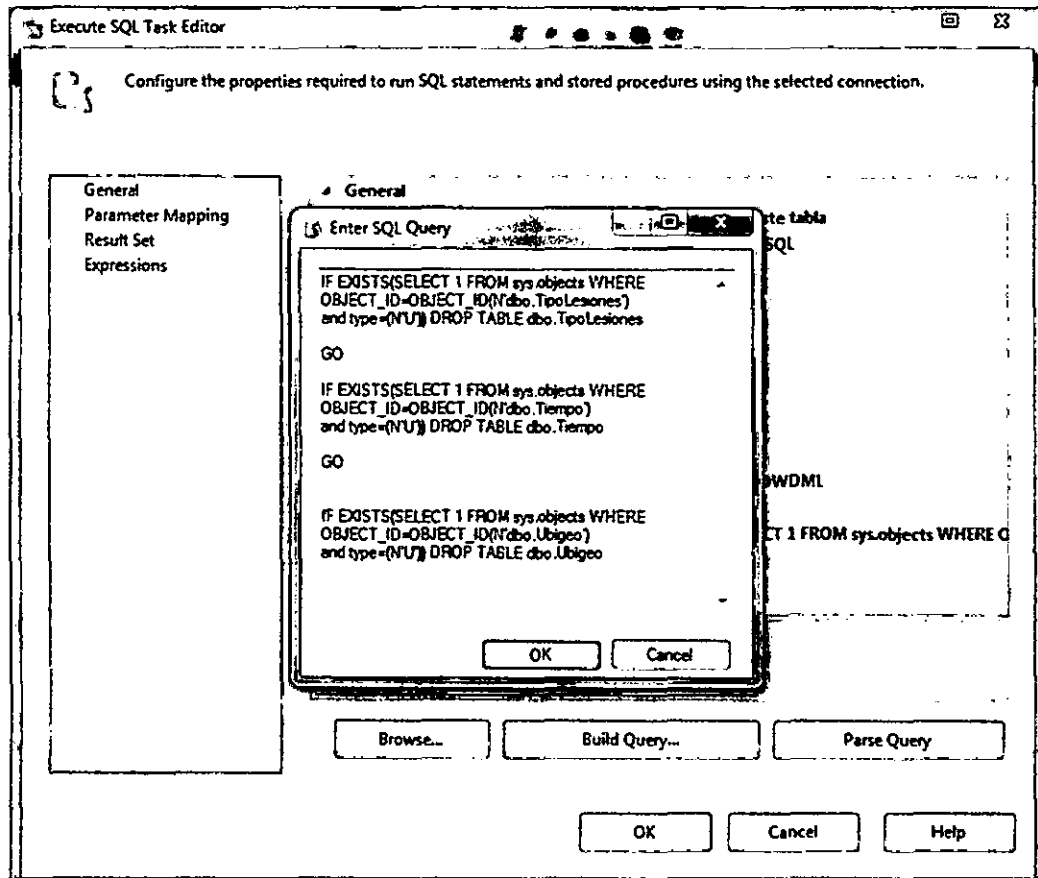
En el paquete (Fig. 4.5) se ha utilizado el control Tarea Ejecutar SQL, a continuación se describe el código SQL incorporado en cada uno de los controles del paquete indicando la funcionalidad que cumple.

Fig. 4.5 Paquete Maestro Lesiones



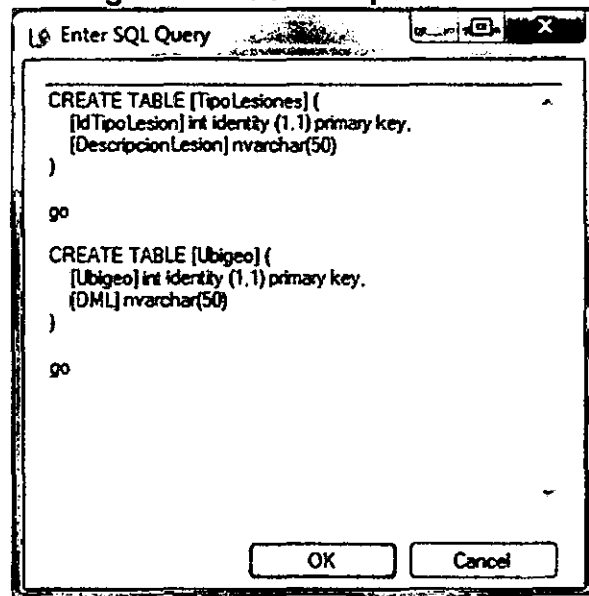
En el control denominado "compara si existe una tabla" se ha implementado código SQL para verificar si existen las estructuras Tipo de Lesiones, Ubicación Geográfica y Tiempo (ver Fig. 4.6).

Fig. 4.6 Verifica Existencia de Estructuras



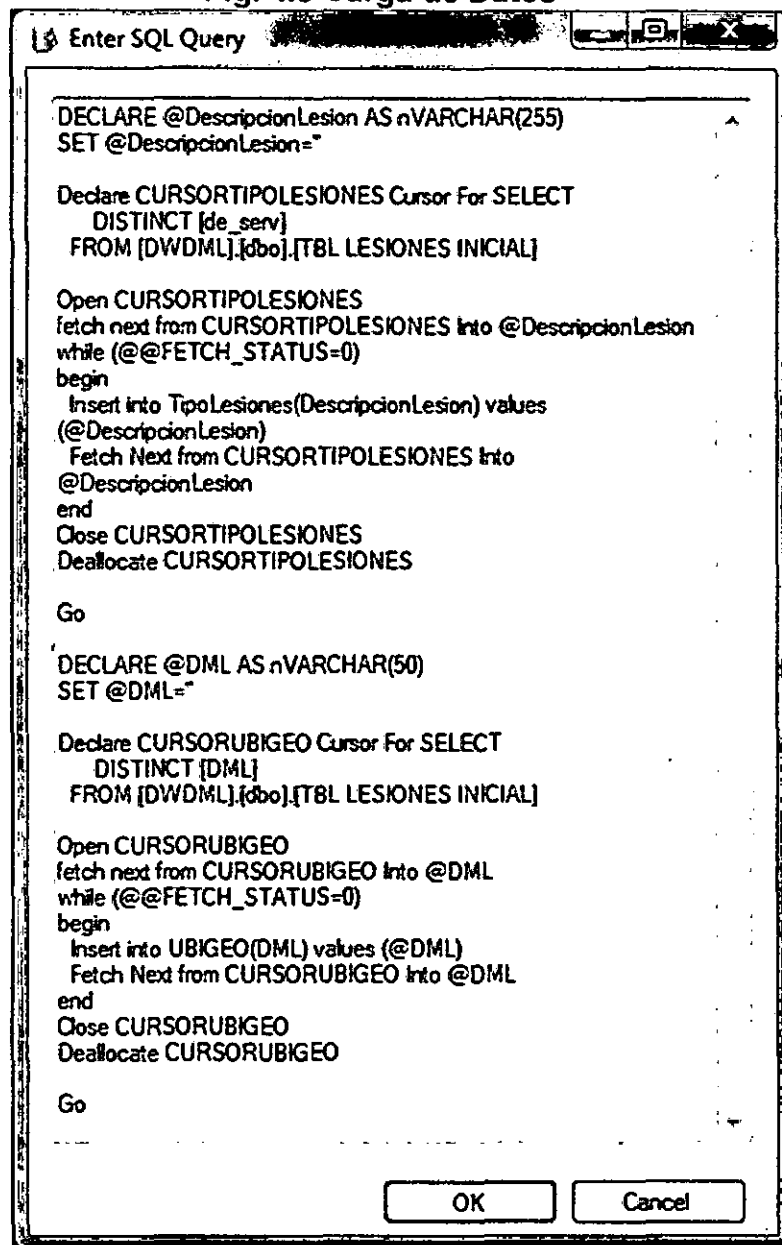
En el control denominado "creación tabla y pk" se ha implementado código SQL para crear la estructura Tipo de Lesiones y se define su clave primaria, del mismo modo la estructura Ubicación Geográfica y se define su clave primaria (ver Fig. 4.7).

Fig. 4.7 Creación TipoLesiones



En el control denominado "llenado de tabla" se ha implementado código SQL para cargar los datos de la estructura [TBL LESIONES INICIAL] a la estructura Tipo de Lesiones utilizando un cursor para su carga de datos, del mismo modo se cargan los datos de la estructura [TBL LESIONES INICIAL] a la estructura Ubicación Geográfica (ver Fig. 4.8).

Fig. 4.8 Carga de Datos



En el control denominado "llenado IdTipoLesion y Ubigeo de la tabla Lesiones" se ha implementado código SQL para crear la estructura Lesiones, verificando previamente su existencia, para finalmente proceder a su carga de datos (ver Fig. 4.9).

Fig. 4.9 Estructura Lesiones

```
IF EXISTS(SELECT 1 FROM sys.objects WHERE
OBJECT_ID=OBJECT_ID(N'dbo.Lesiones')
and type=(N'U')) DROP TABLE dbo.Lesiones

GO

create TABLE dbo.Lesiones(
[Clave] int identity (1,1) primary key,[Ubigeo] int,[IdTipoLesion] int,
[Indice] char(8),[ti_sexo] char(1),[id_tipo_doc] float,[nu_docu_pers]
varchar(50),[ca_edad] int,[de_serv] varchar(100),
[DML] varchar(50)
)

GO

insert into dbo.Lesiones([de_serv],[ti_sexo],[id_tipo_doc],
[nu_docu_pers],[ca_edad],[Indice],[DML]
)
SELECT [de_serv],[ti_sexo],[id_tipo_doc],[nu_docu_pers],
[ca_edad],
RIGHT(RTRIM([fe_ingr_serv]),4)+SUBSTRING
([fe_ingr_serv],4,2)+LEFT(RTRIM([fe_ingr_serv]),2),
[DML]
FROM [DWDML].[dbo].[TBL LESIONES INICIAL]

GO

update [DWDML].[dbo].Lesiones set IdTipoLesion = (select
IdTipoLesion from [DWDML].[dbo].TipoLesiones
where [DWDML].[dbo].Lesiones.[de_serv] like [DWDML].
[dbo].TipoLesiones.DescripcionLesion)

GO

update [DWDML].[dbo].Lesiones set Ubigeo = (select Ubigeo from
[DWDML].[dbo].UBIGEO
where [DWDML].[dbo].Lesiones.[DML] like [DWDML].
[dbo].Ubigeo.DML)

GO

ALTER TABLE [DWDML].[dbo].Lesiones DROP COLUMN [de_serv]

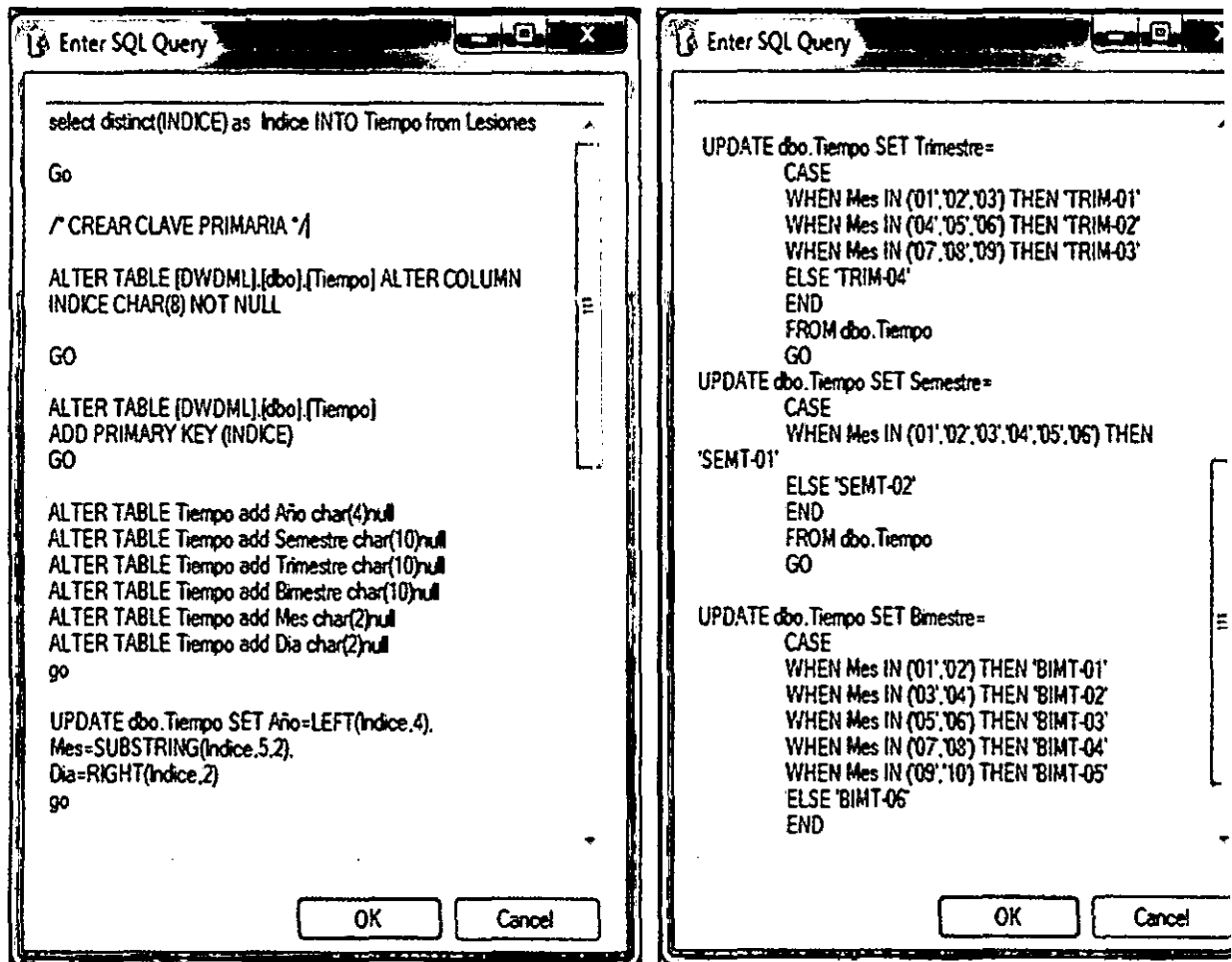
GO

ALTER TABLE [DWDML].[dbo].Lesiones DROP COLUMN [DML]

GO
```

En el control denominado "Creación y carga estructura tiempo" se ha implementado código SQL para crear la estructura tiempo, para luego proceder a su carga de datos (ver Fig. 4.10).

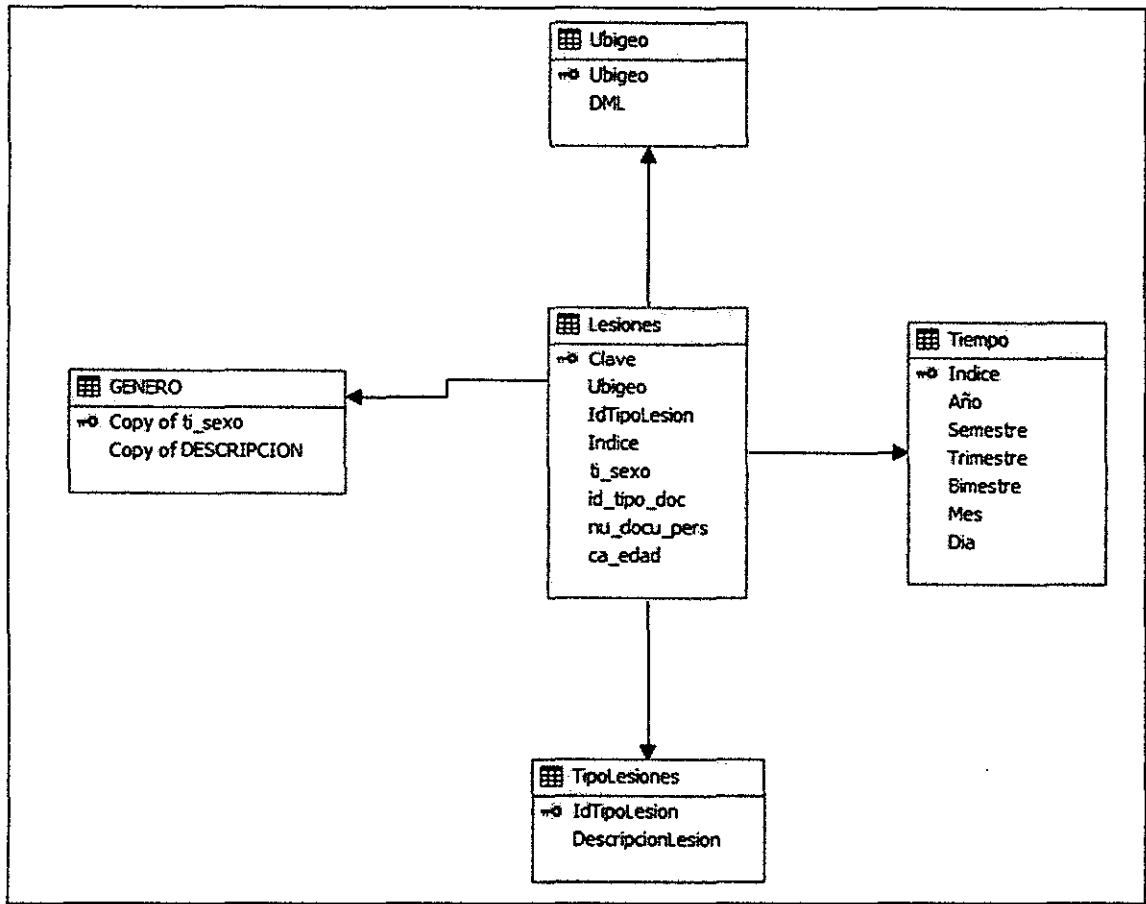
Fig. 4.10 Creación y Carga de Datos Estructura Tiempo



4.4. MODELO DEL DATAWAREHOUSE

Luego de diseñar las estructuras del Datawarehouse de acuerdo a los datos disponibles, y luego de aplicar las herramientas de extracción, transformación y carga se ha obtenido el modelo de Datawarehouse, que se utilizará como fuente de datos para la elaboración de los modelos de minería de datos (Fig. 4.11)

Fig. 4.11 Datawarehouse Lesiones



4.5. MODELOS DE MINERÍA DE DATOS

4.5.1. ALGORITMO DE BAYES NAIVE DE MICROSOFT

Es un algoritmo de clasificación basado en el Teoremas de Bayes y que Microsoft SQL Server Analysis Services proporciona para el modelado de predicción.

Desde el punto de vista computacional, este algoritmo es menos complejo que otros algoritmos de Microsoft y, por tanto, resulta útil para generar rápidamente modelos de minería de datos para descubrir relaciones entre columnas de entrada y columnas de predicción. Se puede utilizar este algoritmo para realizar la exploración inicial de los datos y, más adelante, aplicar los resultados para crear modelos de minería de datos adicionales

con otros algoritmos más complejos y precisos desde el punto de vista computacional.

CASO DE ANÁLISIS

Se planifica un escenario para medir la probabilidad de cada uno de los estados de las columnas o atributos de entrada (Género, Edad y Ubicación Geográfica) para predecir un estado de la columna de predicción (Tipos de Lesiones).

Una vez planificado el modelo se entrena dicho modelo, los resultados se almacenan como un conjunto de patrones que se pueden explorar o utilizar para realizar predicciones. Puede crear consultas para devolver las predicciones sobre cómo se relacionan los nuevos datos con el atributo de predicción, o puede recuperar estadísticas que describan las correlaciones que ha hallado el modelo.

FUNCIONAMIENTO DEL ALGORITMO

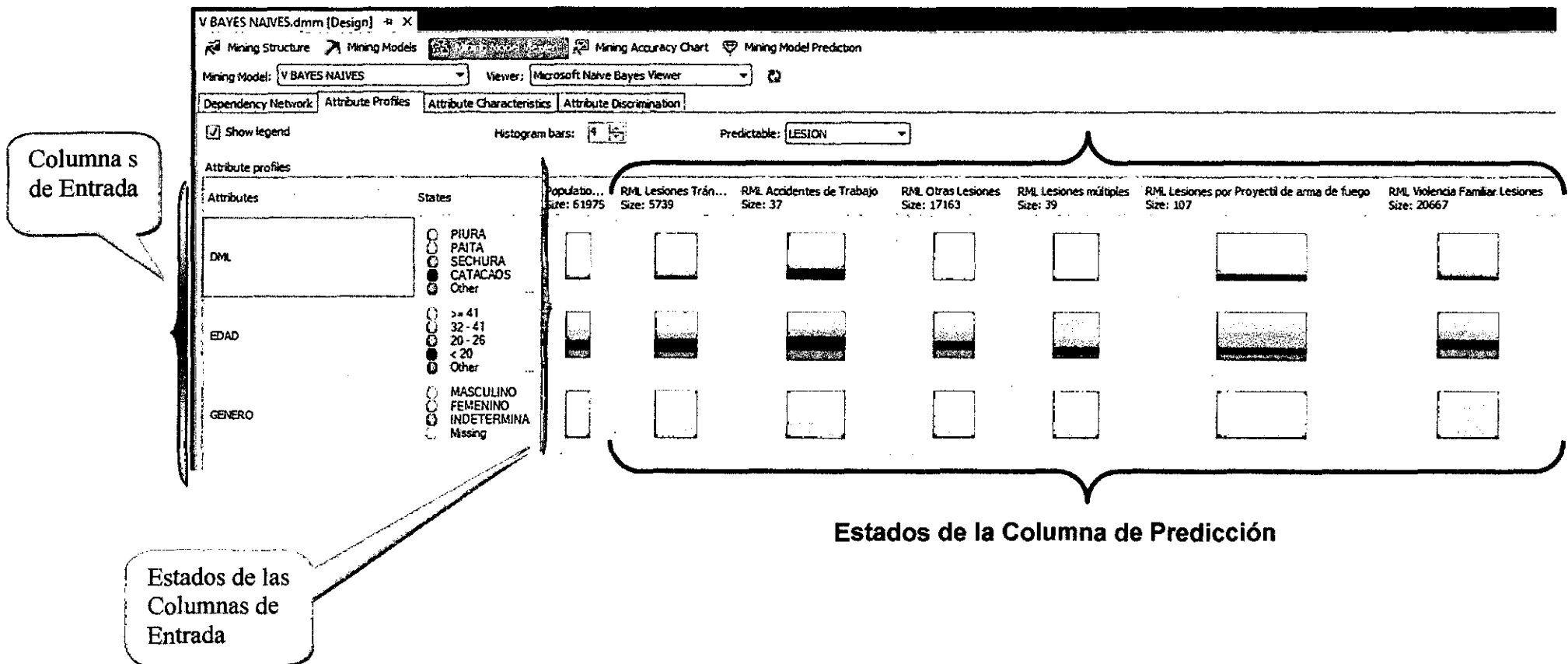
El algoritmo de Bayes Naive de Microsoft calcula la probabilidad de cada estado de cada columna de entrada (Género, Edad y Ubicación Geográfica), dado cada posible estado de la columna de predicción (Tipos de Lesiones).

Para comprender cómo funciona, se utiliza el visor de Bayes Naive de Microsoft (Fig. 4.12) para consultar una representación visual del modo en que el algoritmo distribuye los estados.

En la Fig. 4.12 el visor Bayes naive de Microsoft muestra cada columna de entrada (Género, Edad y Ubicación Geográfica) del conjunto de datos e indica cómo se distribuyen los estados de cada columna, dado cada estado de la columna de predicción (Tipo de Lesiones).

Esta vista del modelo se utilizaría para identificar las columnas de entrada que son importantes para diferenciar los distintos estados de la columna de predicción.

Fig. 4.12 Distribución de los Estados Entrada y Predicción



DESCRIPCIÓN DEL MODELO DE BAYES NAIVE E INTERPRETACIÓN DE RESULTADOS

El visor (Fig 4.12) proporciona colores y valores porcentuales para las distribuciones, de modo que se puede ver la distribución porcentual de los estados de entrada (Género, Edad, Ubicación Geográfica) después de haber seleccionado un estado de la columna de predicción (Tipos de Lesiones). Del mismo modo se describe el modelo creado por el algoritmo de Bayes Naive, Se pueden visualizar los estados del atributo de predicción - Tipos de Lesiones los cuales son:

- Lesiones por proyectil arma de fuego
- Lesiones por suceso de tránsito
- Lesiones por mordedura canina
- Lesiones por violencia familiar
- Lesiones por arma blanca
- Lesiones por accidente de trabajo
- Lesiones múltiples
- Otras lesiones

Asimismo se puede ver la correlación con los estados de los diferentes atributos de entrada:

Género (Masculino, Femenino)

Edad (Mayor o igual a 41 años, Entre 32 a 41 años inclusive, Entre 26 a 32 años inclusive, Entre 20 a 26 años inclusive, Menor a 20 años)

Ubicación Geográfica (Piura, Paita, Sechura, Catacaos), a partir de este modelo se pueden hacer interpretaciones para posteriormente tomar decisiones. Entre las múltiples interpretaciones que se pueden hacer a continuación describimos las lesiones por violencia familiar:

- a) Con el atributo de entrada Género podemos observar la distribución de las lesiones por violencia familiar (Fig. 4.12), donde se observa que el mayor porcentaje de lesiones por

violencia familiar ocurre sobre las mujeres con el 81.1% de los casos registrados, y el menor porcentaje de los casos por este mismo tipo de hechos ocurre sobre los hombres con el 18.8%.

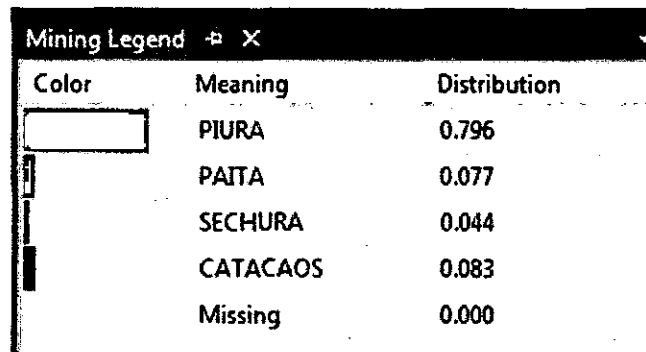
Color	Meaning	Distribution
	MASCULINO	0.188
	FEMENINO	0.811
	INDETERMINADO	0.000
	Missing	0.000

- b) Con el atributo de entrada Edad podemos observar la distribución de las lesiones por violencia familiar (Fig. 4.12), donde se observa que el mayor porcentaje de lesiones por violencia familiar ocurre sobre las personas con edades comprendidas entre 32 a 41 años con el 23.7% de los casos registrados, y el menor porcentaje de los casos por este mismo tipo de hechos ocurre sobre las personas que tienen edad superior a los 41 años con el 17.1%.

Color	Meaning	Distribution
	>= 41	0.171
	32 - 41	0.237
	20 - 26	0.188
	< 20	0.194
	26 - 32	0.210
	Missing	0.000

- c) Con el atributo de entrada Ubicación Geográfica (DML) podemos observar la distribución de las lesiones por violencia familiar (Fig. 4.12), donde se observa que el mayor porcentaje de lesiones por violencia familiar ocurre sobre las personas que viven en la zona de Piura con el 79.6% de los casos registrados, y el menor porcentaje de los casos por este mismo tipo de hechos ocurre

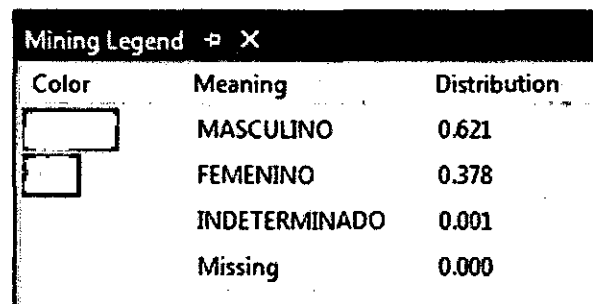
sobre las personas que viven en la zona de Sechura con el 4.4%.



Color	Meaning	Distribution
	PIURA	0.796
	PAITA	0.077
	SECHURA	0.044
	CATACAOS	0.083
	Missing	0.000

Como segundo caso a analizar describimos las lesiones registradas por suceso de tránsito.





- a) Con el atributo de entrada Género podemos observar la distribución de las lesiones por suceso de tránsito (Fig. 4.12), donde se observa que el mayor porcentaje de lesiones por suceso de tránsito ocurre sobre los hombres con el 62.1% de los casos registrados, y el menor porcentaje de los casos por este mismo tipo de hechos ocurre sobre las mujeres con el 37.8%.




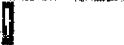


Color	Meaning	Distribution
	MASCULINO	0.621
	FEMENINO	0.378
	INDETERMINADO	0.001
	Missing	0.000

- b) Con el atributo de entrada Edad podemos observar la distribución de las lesiones por suceso de tránsito (Fig. 4.12), donde se observa que el mayor porcentaje de lesiones por suceso de tránsito ocurre sobre las personas con edad superior a los 41 años con el 28.8% de los casos registrados, y el menor porcentaje de los casos por este mismo tipo de hechos ocurre

sobre las personas que tienen edad comprendida entre 26 a 32 años con el 14.3%.

Mining Legend		
Color	Meaning	Distribution
	>= 41	0.288
	32 - 41	0.149
	20 - 26	0.152
	< 20	0.267
	26 - 32	0.143
	Missing	0.000

- c) Con el atributo de entrada Ubicación Geográfica (DML) podemos observar la distribución de las lesiones por suceso de tránsito (Fig. 4.12), donde se observa que el mayor porcentaje de lesiones por suceso de tránsito ocurre sobre las personas que viven en la zona de Piura con el 79% de los casos registrados, y el menor porcentaje de los casos por este mismo tipo de hechos ocurre sobre las personas que viven en la zona de Sechura con el 5%.

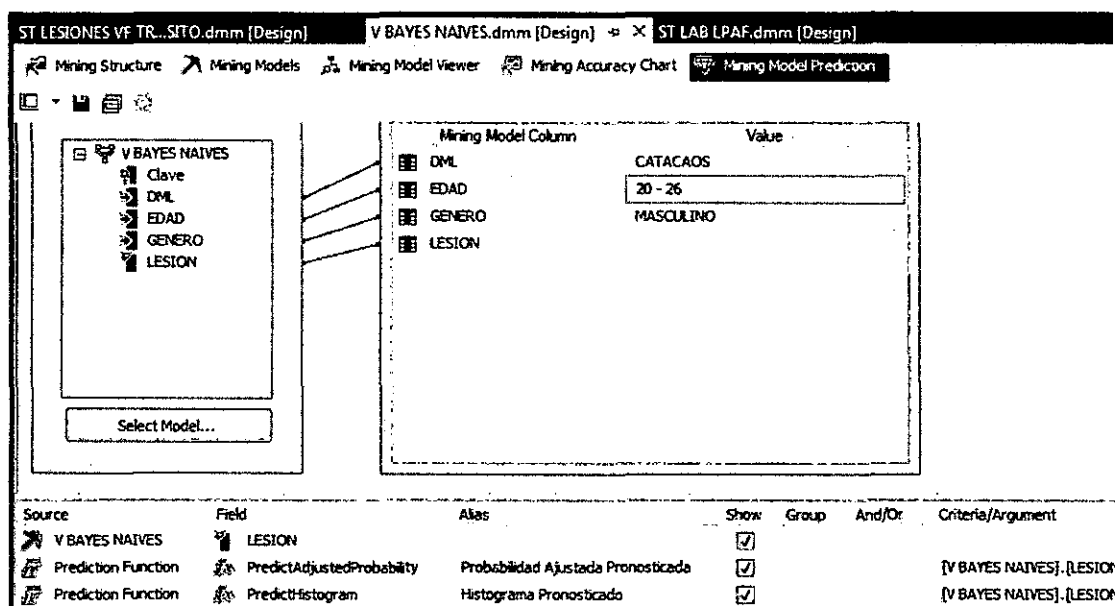
Mining Legend		
Color	Meaning	Distribution
	PIURA	0.790
	PAITA	0.077
	SECHURA	0.050
	CATACAOS	0.083
	Missing	0.000

CASO DE PREDICCIÓN MEDIANTE CONSULTAS SINGLETON

Con la herramienta de Analysis Service también podemos predecir mediante consultas singleton de acuerdo a los estados de las columnas de entrada (Edad, Género y Ubicación Geográfica) cuál será la probabilidad de que se produzca uno de los estados de la

columna de predicción (Tipos de Lesiones). En el siguiente caso se desea predecir el histograma para los diferentes estados de la columna de predicción (Tipos de Lesiones) con los siguientes parámetros, que el suceso se haya registrado en Catacaos, que la persona lesionada tenga entre 20 a 26 años de edad inclusive y que sea de sexo Masculino.

Fig. 4.13 - Consulta Singleton de Predicción



Después de ejecutar la consulta singleton descrita, se obtiene el siguiente resultado (Fig. 4.14):

Fig. 4.14 - Resultado Consulta

ST LESIONES VF TR...SITO.dmm [Design]		V BAYES NAIVES
Mining Structure Mining Models Mining Model Viewer		
LESION Probabilidad Ajustada Pronosticada		
RML Violencia Familiar Lesiones 0.0505813324843352		

Donde podemos observar que de acuerdo a los parámetros de la consulta las lesiones por violencia familiar son las que tienen una mayor probabilidad de ocurrir.

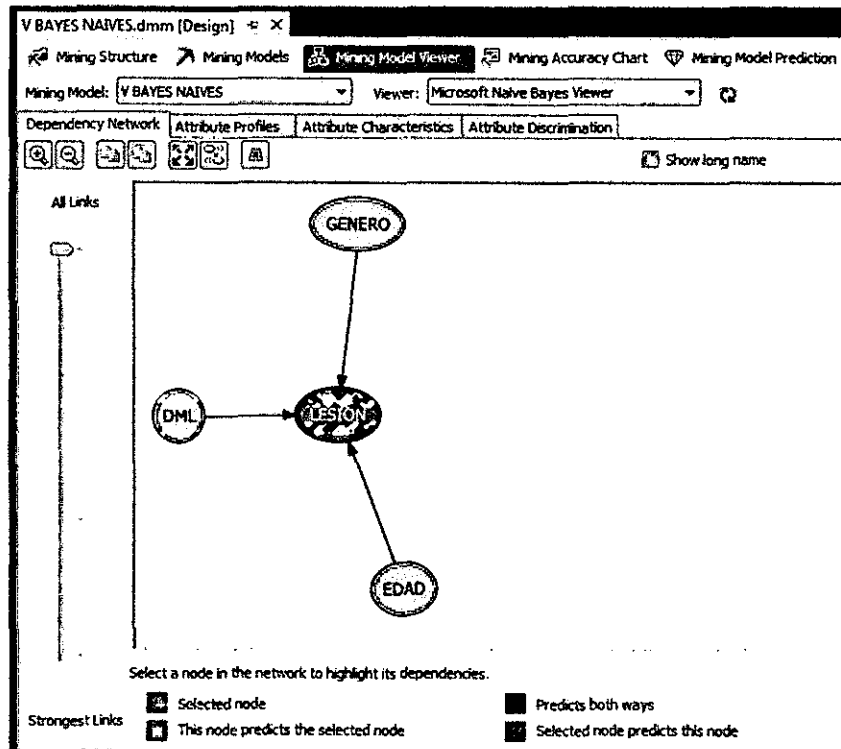
Del mismo modo podemos observar en el siguiente resultado (Fig. 4.15), el orden probabilístico predicho en el que podría ocurrir un caso de lesiones de acuerdo a los parámetros ingresados en la consulta singleton sobre el modelo elaborado, observando que en primer lugar tenemos las lesiones por violencia familiar, en segundo lugar las lesiones de tránsito y en tercer lugar las lesiones por arma blanca.

Fig. 4.15 - Histograma Predicho

Histograma Pronosticado			
- Histograma Pronosticado			
LESION	\$SUPPORT	\$PROBABILITY	\$ADJUSTEDPROBABILITY
RML Violencia Familiar Lesiones	31230.1124678447	0.503914682821214	0.0505813324843352
RML Lesiones Tránsito	23171.4221024077	0.373883373979955	0.215571252551613
RML Lesiones por arma blanca	4523.126107406	0.0729830755531424	0.0720405348991419
RML Lesiones por Proyectoil de arma de fuego	2384.62235575901	0.038477165885583	0.0381022372636482
RML Accidentes de Trabajo	449.971176303571	0.00726052724975508	0.0072359984641204
RML Lesiones por Mordedura Canina	211.746435596359	0.00341664276879967	0.00336479895167326
RML Lesiones múltiples	0.999838670646124	1.61329353875938E-05	1.60754904891886E-05
RML Lesiones	0.999838670646124	1.61329353875938E-05	2.33414415719219E-06
RML Otras Lesiones	0.999838670646124	1.61329353875938E-05	2.56874433631291E-06
	0.999838670646124	1.61329353875938E-05	1.61329353875938E-05

Asimismo, podemos visualizar la red de dependencias entre las columnas de entrada (Género, Edad y Ubicación Geográfica) y la columna de predicción (Tipos de Lesiones) (Fig. 4.16)

Fig.4.16 - Red de Dependencias



En la Fig. 4.16, el nodo seleccionado Lesión (Tipo de Lesiones) se puede predecir por los nodos DML (Ubicación Geográfica), Edad y Género.

DATOS REQUERIDOS PARA LOS MODELOS BAYES NAIVE

Al preparar los datos para su uso en un modelo de entrenamiento Bayes naive, conviene comprender qué requisitos son imprescindibles para el algoritmo, incluidos el volumen de datos necesario y la forma en que estos datos se utilizan. Los requisitos para un modelo Bayes naive son los siguientes:

- **Una columna de una sola clave:** cada modelo debe contener una columna numérica o de texto que identifique cada registro de manera única. No están permitidas las claves compuestas. En nuestro modelo la columna se denomina clave y es de tipo de datos numérico entero.

- **Columnas de entrada:** en un modelo Bayes naive, todas las columnas deben ser discretas o discretizadas.

En un modelo Bayes naive, también es importante asegurarse de que los atributos de entrada sean independientes unos de otros. Esto es particularmente importante al utilizar el modelo para la predicción. El motivo es que, si utiliza dos columnas de datos que ya están estrechamente relacionadas, el efecto sería multiplicar la influencia de esas columnas, lo que puede ocultar otros factores que influyen en el resultado.

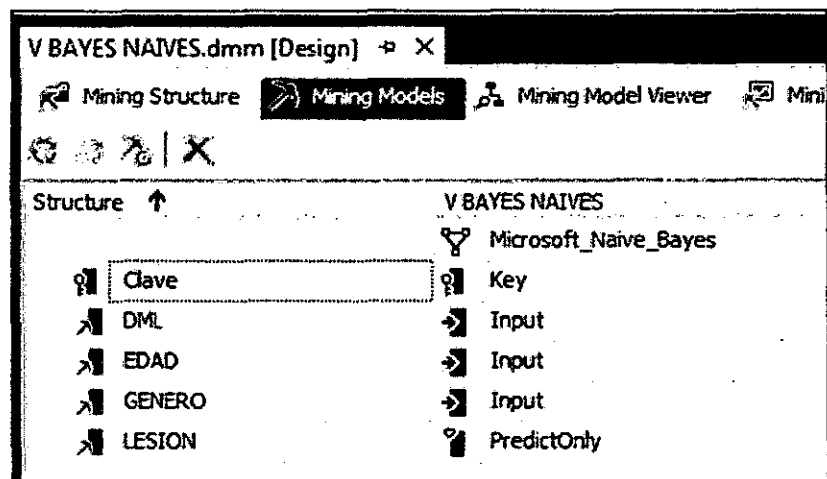
En nuestro modelo las columnas de entrada se denominan Género de tipo de dato discreto, Edad de tipo de dato discretizada y Ubicación Geográfica de tipo de dato discreto.

- **Al menos una columna de predicción:** el atributo de predicción debe contener valores discretos o discretizados.

En nuestro modelo la columna de predicción se denomina Lesión (donde los estados son los diferentes tipos de lesión) de tipo de dato discreto.

Los valores de la columna predecible se pueden tratar como entradas. Esto puede ser útil si explora un nuevo conjunto de datos, para encontrar relaciones entre las columnas.

Fig.4.17 - Modelo Bayes Naive



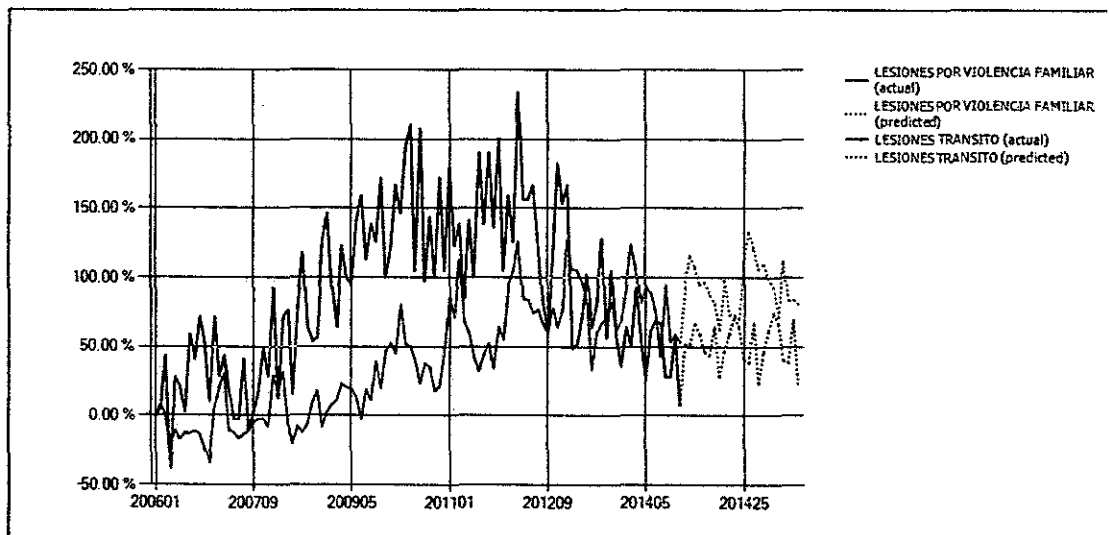
Finalmente un modelo Bayes naive no admite tipos de datos continuos, todos los valores de las columnas de entrada se tratan como discretos o discretizados.

4.5.2. ALGORITMO DE SERIE TEMPORAL

El Pronóstico de series de tiempo asume que los datos del pasado pueden ayudar a explicar los valores futuros. Es importante saber que en algunas situaciones, puede haber circunstancias que no se reflejan en los datos históricos. Por ejemplo, el Gobierno Peruano podría ejercer un mayor control sobre el tránsito vehicular lo que podría ocasionar que el número de lesiones producidas por sucesos de tránsito disminuya en el futuro. En este tipo de situaciones, un pronóstico de series de tiempo no puede ser el mejor enfoque o no debe ser el único criterio considerado. A menudo, los diferentes enfoques de previsión se combinan para proporcionar las predicciones más precisas.

Una serie de tiempo es un conjunto de valores observados durante un período de tiempo, por lo general a intervalos regulares (semanal, trimestral, mensual, anual). Los datos de Lesiones de Tránsito y Lesiones por Violencia Familiar se presentan en un formato gráfico, con el intervalo de tiempo a lo largo del eje x y los valores a lo largo del eje y, como muestra la Figura 4.18

Fig.4.18 - Modelo Series de Tiempo



En cuanto a la comprensión de cómo se modifica un valor de un período a otro, y cómo predecir los valores futuros, los datos de series de tiempo tiene varias características principales:

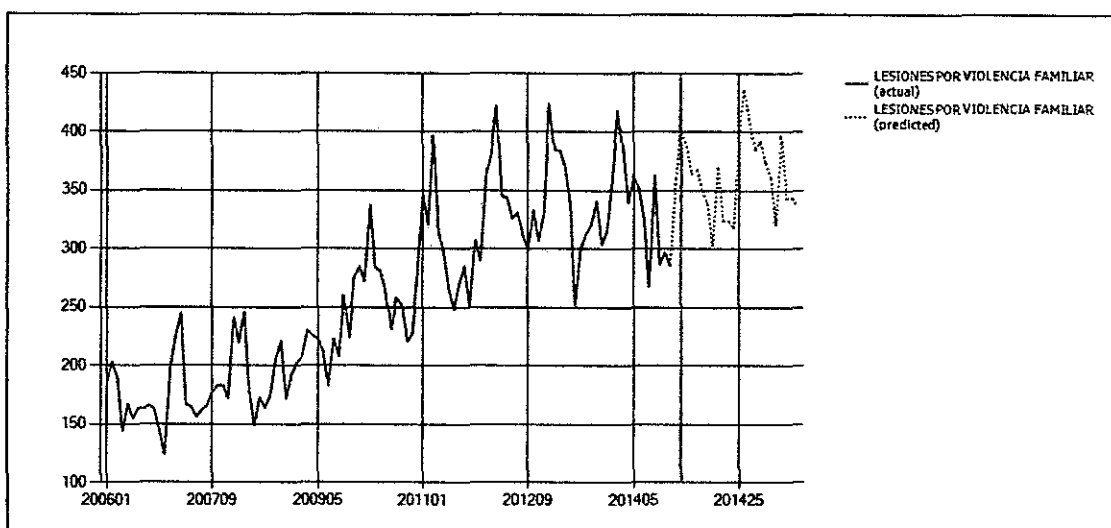
- **Nivel de base.** El nivel de base se define típicamente como el valor medio de la serie. En algunos modelos de predicción, el nivel de base se define como el valor inicial de los datos de series. En nuestro caso el nivel de base corresponde a la cantidad de lesiones tanto por violencia familiar como por suceso de tránsito del mes de enero del año 2006. Lo descrito se aprecia en la Fig. 4.18.
- **Tendencia.** Una tendencia se define generalmente como la forma en que la serie cambia de un período a otro.

Una manera de identificar un valor de base y la tendencia es aplicar una técnica de regresión. El término regresión significa estudiar la relación entre las variables. En este caso, es la relación entre la variable independiente del tiempo y la variable dependiente del número de lesiones de tránsito o número de lesiones de violencia familiar.

CASO DE ANÁLISIS

Se planifica un escenario para pronosticar y encontrar patrones de comportamiento en los tipos de lesiones (lesiones por violencia familiar, lesiones por suceso de tránsito, lesiones por proyectil de arma de fuego, lesiones por arma blanca) más frecuentes e importantes que se presentan diariamente en la División Médico Legal II de Piura, Sechura, Paita y Catacaos.

Fig.4.19 - Modelo Series de Tiempo
Lesiones por Violencia Familiar



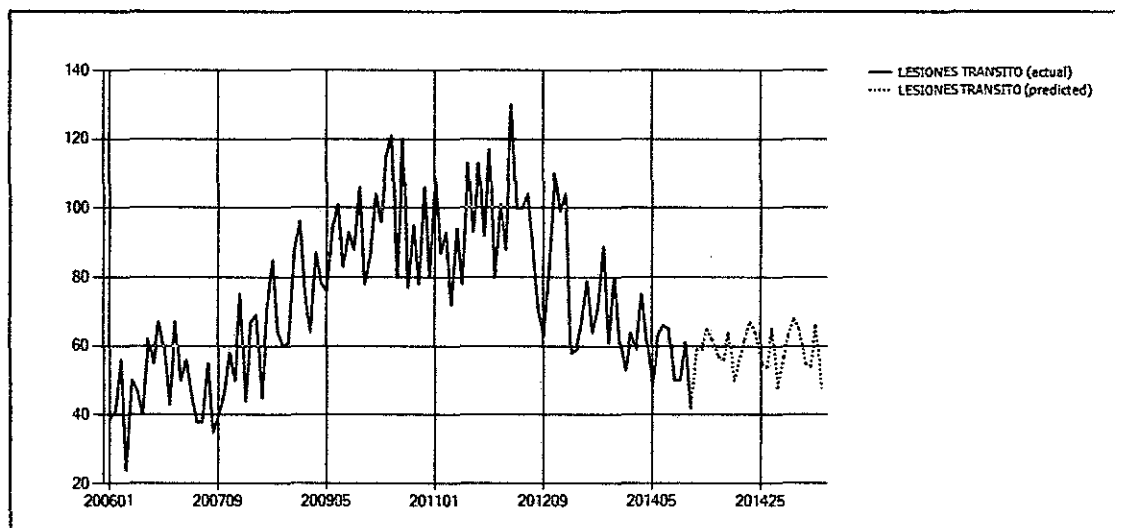
INTERPRETACIÓN

De acuerdo al modelo obtenido se han analizado datos desde el mes de enero del año 2006 hasta el mes de diciembre del 2014, encontrando patrones de comportamiento los cuales indican que el mayor número de casos de lesiones de violencia familiar se registra entre los meses de enero a marzo de cada año y por el contrario la menor cantidad de casos por este mismo hecho se registra entre los meses junio a setiembre de cada año. Del mismo modo el modelo predice que para el mes de febrero del año 2016 (201426) se registrará una cantidad de 436 casos de lesiones por violencia familiar (el número más alto para ese año) y para el mes de agosto del año 2016 (201432) se predice (el número más bajo para ese año) una cantidad de 321 casos. Para

un mayor detalle observar la siguiente tabla con los resultados predichos del modelo para los próximos meses.

ST LESIONES VF TR...SITO.dmm [Design]		V BAYES NAIVES.dmm [Design]
Mining Structure		Mining Models
Mining Model Viewer		Mining
LESIONES POR ...		Expression
360	-	Expression
		LESIONES POR VIOLENCIA FAMILIAR
\$TIME		
201417		368
201418		350
201419		339
201420		303
201421		370
201422		324
201423		324
201424		318
201425		393
201426		436
201427		412
201428		385
201429		392
201430		374
201431		361
201432		321
201433		397
201434		343
201435		344
201436		338

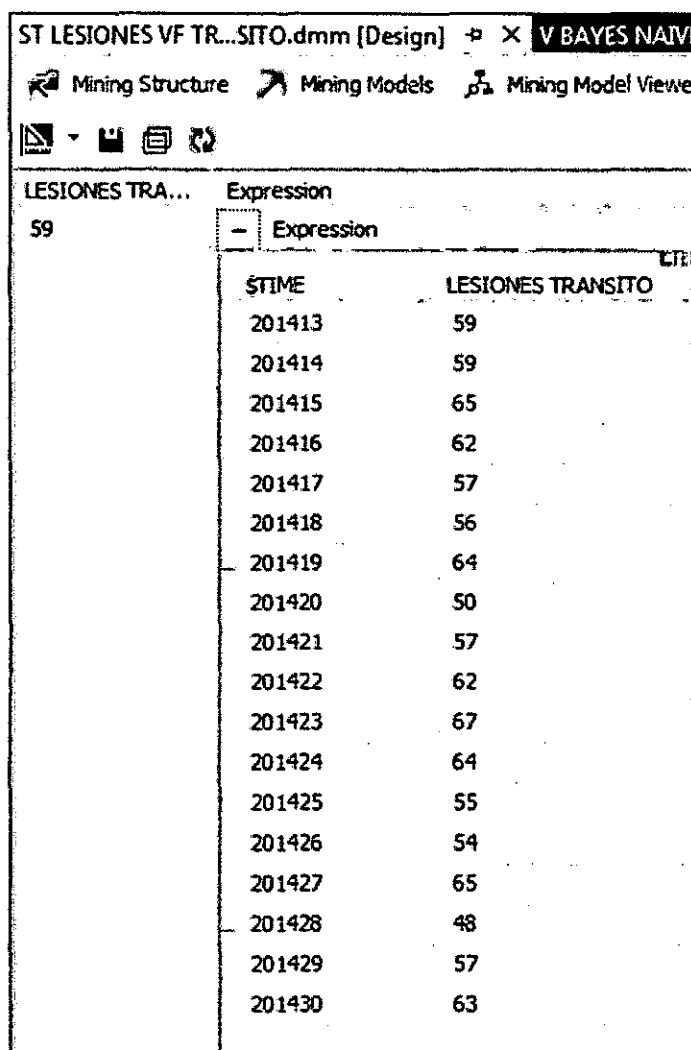
Fig.4.20 - Modelo Series de Tiempo
Lesiones por suceso de tránsito



INTERPRETACIÓN

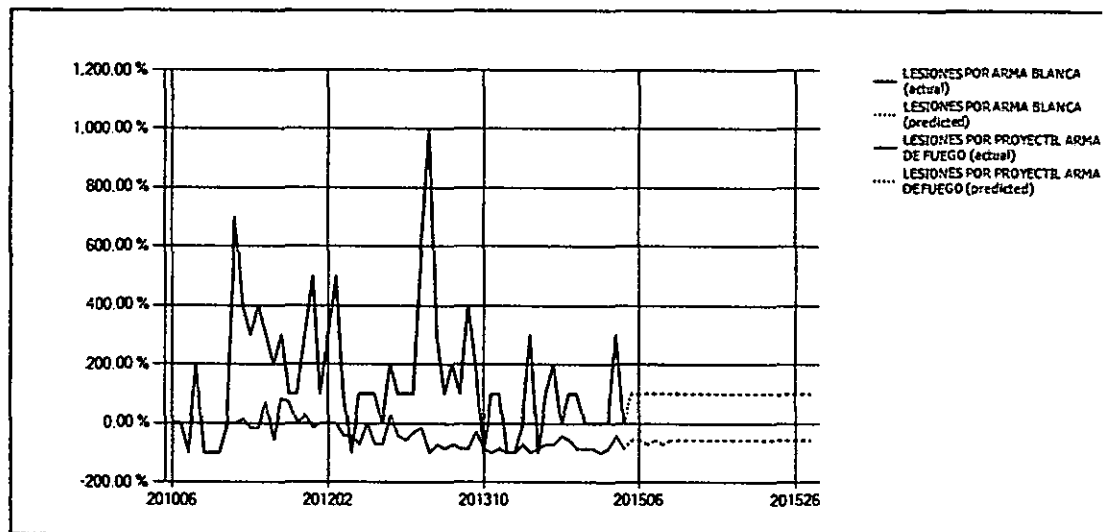
De acuerdo al modelo obtenido se han analizado datos desde el mes de enero del año 2006 hasta el mes de diciembre del 2014, encontrando patrones de comportamiento los cuales indican que el número de casos por lesiones de suceso de tránsito se han ido incrementando en el tiempo desde el mes de enero del año 2006 encontrando el mayor registró de estos casos en el mes de marzo del año 2012, a partir de ese momento la cantidad de casos registrados ha disminuido considerablemente. Del mismo modo el modelo predice que para los próximos meses la cantidad de casos presentará un incremento mínimo, lo descrito se puede visualizar en la siguiente tabla (Fig. 4.21).

Fig.4.21 - Predicción Lesiones por suceso de Tránsito



Expression	LESIONES TRANSITO
59	
201413	59
201414	59
201415	65
201416	62
201417	57
201418	56
201419	64
201420	50
201421	57
201422	62
201423	67
201424	64
201425	55
201426	54
201427	65
201428	48
201429	57
201430	63

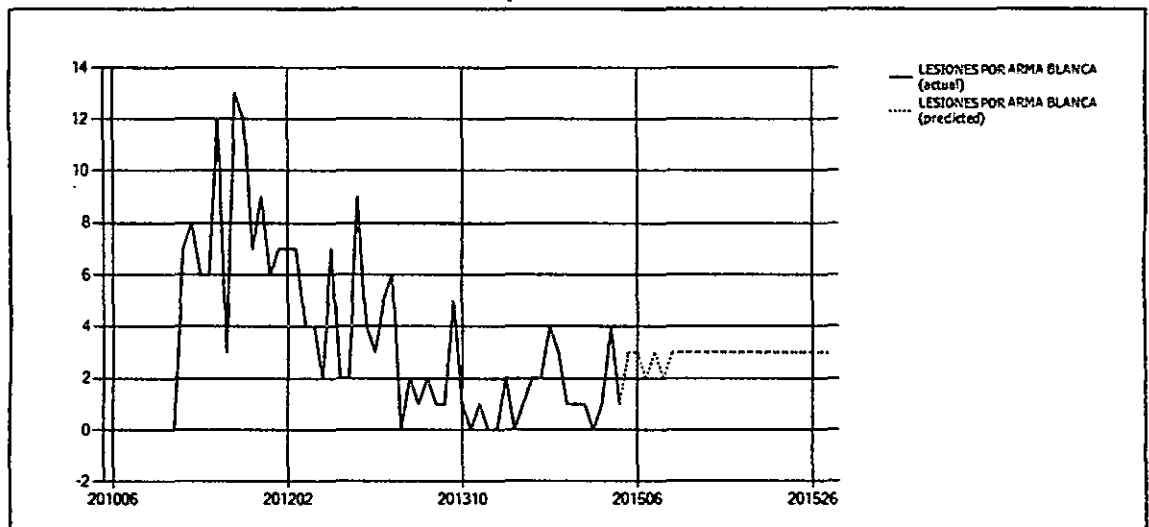
Fig.4.22 - Modelo Series de Tiempo
Lesiones por arma blanca y arma de fuego



INTERPRETACIÓN

De acuerdo al modelo obtenido se han analizado datos desde el mes de junio del año 2010 hasta el mes de abril del año 2015, encontrando patrones de comportamiento los cuales indican que el número de casos de lesiones por proyectil de arma de fuego es superior al número de casos de lesiones por arma blanca, además el registro de lesiones por proyectil de arma de fuego presenta un comportamiento poco predecible en relación al registro de lesiones por arma blanca que presenta un comportamiento más predecible a lo largo de los años analizados.

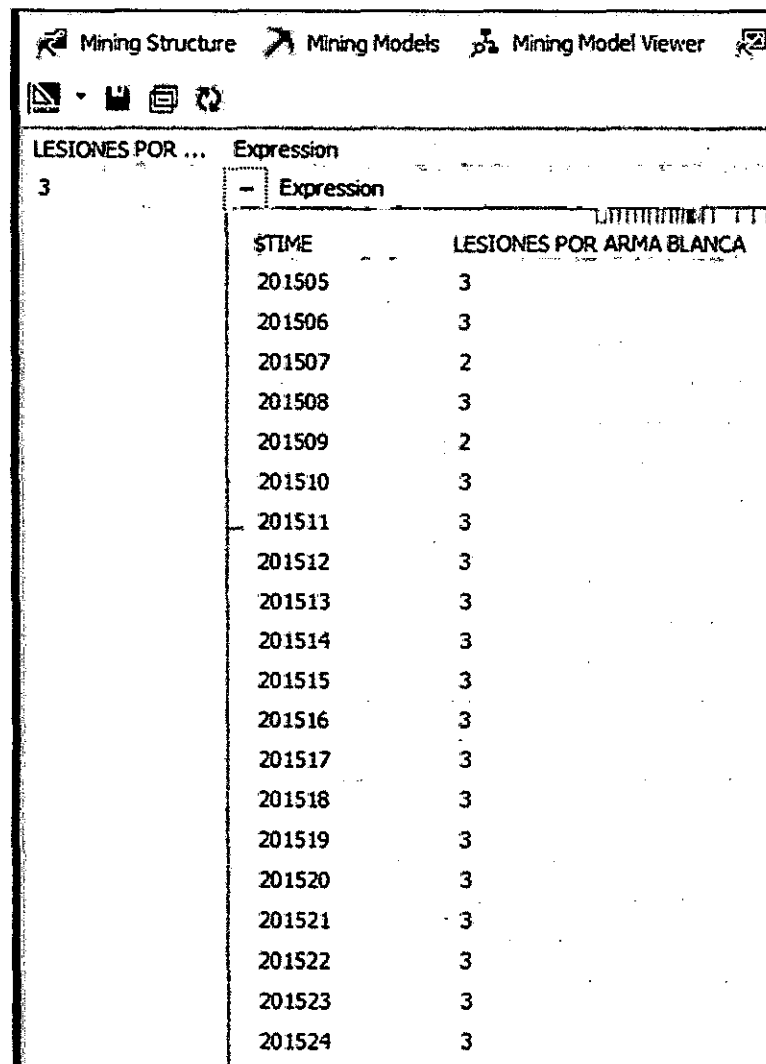
Fig.4.23 - Modelo Series de Tiempo
Lesiones por arma blanca



INTERPRETACIÓN

De acuerdo al modelo obtenido se han analizado datos desde el mes de junio del año 2010 hasta el mes de abril del año 2015, encontrando patrones de comportamiento los cuales indican que el número de casos de lesiones por arma blanca ha ido disminuyendo en el tiempo y se predice que continuará así a lo largo de los próximos 20 meses, registrando una cantidad casi constante, tal como se muestra en la consulta DMX obtenida del modelo que aparece en la siguiente tabla.

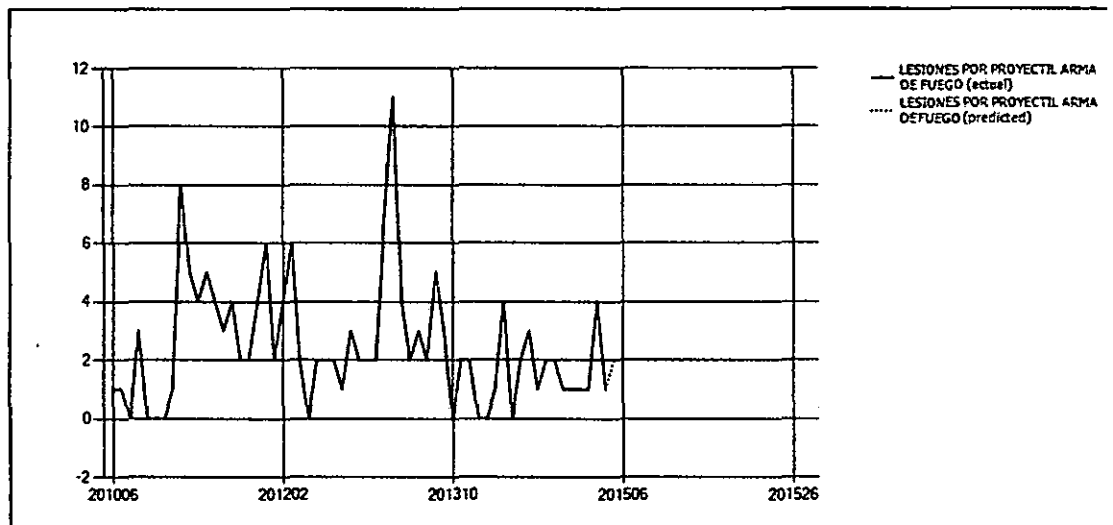
Fig.4.24 - Predicción lesiones por arma blanca



The screenshot shows the Mining Model Viewer interface. The main window displays a table with two columns: '\$TIME' and 'LESIONES POR ARMA BLANCA'. The table contains 20 rows of data, representing predictions for the next 20 months starting from 201505. The values for 'LESIONES POR ARMA BLANCA' are mostly 3, with a few instances of 2 (at 201507 and 201509). The interface includes a toolbar with icons for Mining Structure, Mining Models, and Mining Model Viewer, and a menu bar with options like File, Edit, and View.

\$TIME	LESIONES POR ARMA BLANCA
201505	3
201506	3
201507	2
201508	3
201509	2
201510	3
201511	3
201512	3
201513	3
201514	3
201515	3
201516	3
201517	3
201518	3
201519	3
201520	3
201521	3
201522	3
201523	3
201524	3

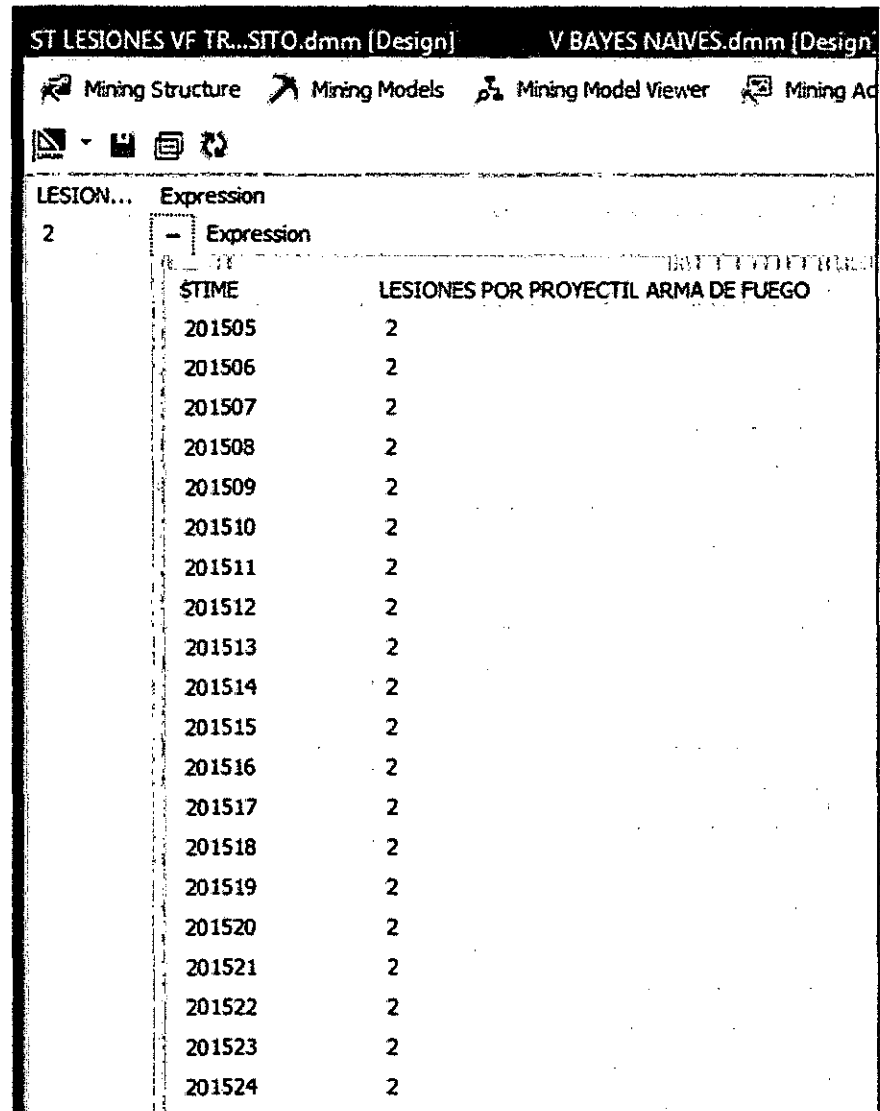
Fig.4.25 - Modelo Series de Tiempo
Lesiones por proyectil arma de fuego



INTERPRETACIÓN

De acuerdo al modelo obtenido se han analizado datos desde el mes de junio del año 2010 hasta el mes de abril del año 2015, encontrando patrones de comportamiento los cuales indican que el número de casos de lesiones por proyectil arma de fuego ha disminuido levemente en el tiempo y se predice que el número de casos presentados durante los próximos 20 meses por este tipo de lesión se mantendrá casi constante tal como se muestra en la consulta DMX obtenida del modelo que aparece en la siguiente tabla (Fig 4.26).

Fig.4.26 - Predicción Lesiones por proyectil de arma de fuego



LESION...	Expression
2	- Expression
	201505 2
	201506 2
	201507 2
	201508 2
	201509 2
	201510 2
	201511 2
	201512 2
	201513 2
	201514 2
	201515 2
	201516 2
	201517 2
	201518 2
	201519 2
	201520 2
	201521 2
	201522 2
	201523 2
	201524 2

DATOS REQUERIDOS PARA LOS MODELOS DE SERIE TEMPORAL

El algoritmo de serie temporal de Microsoft incluye dos algoritmos independientes para analizar series de tiempo:

- ✓ El algoritmo ARTXP, se ha optimizado para predecir el siguiente valor probable de una serie.
- ✓ El algoritmo ARIMA se ha agregado para mejorar la exactitud de la predicción a largo plazo.

De forma predeterminada, Analysis Services utiliza cada algoritmo por separado para entrenar el modelo y, a continuación, combina los

resultados para obtener la mejor predicción de un número variable de predicciones. También puede decidir utilizar solo uno de los algoritmos, dependiendo de los datos y los requisitos de la predicción.

Cada modelo de previsión debe contener una serie de casos, que es la columna que especifica los intervalos de tiempo (en nuestro caso se utilizan intervalos de tiempo mensuales) u otras series sobre las que se produce el cambio. Por ejemplo, los datos del anterior diagrama (Fig. 4.18) muestran las series correspondientes al historial y a la previsión de lesiones por violencia familiar y lesiones por suceso de tránsito para un período de varios meses. Para este modelo, cada tipo de lesión es una serie y la columna de fecha contiene la serie temporal, que también es la serie de casos. Un modelo de serie temporal debe siempre utilizar una fecha, una hora o algún otro valor numérico único para su serie de escenarios.

Los requisitos para un modelo de serie temporal son los siguientes:

Una única columna Key Time

Cada modelo debe contener una columna numérica o de fecha que se utilizará como serie de casos y que define los intervalos de tiempo que utilizará el modelo. El tipo de datos para la columna de clave temporal puede ser un tipo de datos datetime o bien numérico. Sin embargo, la columna debe contener valores continuos y éstos deben ser únicos para cada serie. La serie de casos para un modelo de serie temporal no pueden estar almacenada en dos columnas como por ejemplo una columna Año y una columna Mes.

Una columna predecible

Cada modelo debe contener por lo menos una columna predecible alrededor de la que el algoritmo generará el modelo de serie temporal (en el modelo desarrollado fig. 4.18 del presente proyecto de tesis tenemos las columnas predecibles lesiones por violencia familiar y lesiones por suceso de tránsito, en el otro

modelo fig. 4.22 tenemos las columnas predecibles lesiones por arma blanca y lesiones por proyectil arma de fuego). El tipo de datos de la columna predecible debe contener valores continuos. Por ejemplo, es posible predecir la manera en que los atributos numéricos tales como lesiones por violencia familiar, lesiones por suceso de tránsito, lesiones por arma blanca y lesiones por proyectil arma de fuego varían con el tiempo. Sin embargo, no es posible utilizar como columna predecible una columna que contenga valores discretos tales como el género de las personas que sufren lesiones por violencia familiar o la edad de las personas.

Una columna de clave de serie opcional

Cada modelo puede tener una columna de clave adicional que contenga valores únicos que identifiquen a una serie. La columna de clave de serie opcional debe contener valores únicos. Por ejemplo, un solo modelo puede contener lesiones de varios tipos, siempre y cuando haya un solo registro para cada lesión para cada intervalo de tiempo.

4.5.3 ALGORITMO DE ÁRBOLES DE DECISIÓN

El algoritmo de árboles de decisión es un algoritmo de clasificación y regresión para el modelado de predicción de atributos discretos (Tipos de Lesiones) y continuos. Para los atributos discretos, el algoritmo hace predicciones basándose en las relaciones entre las columnas de entrada (Edad, Género, Ubicación Geográfica) de un conjunto de datos. Utiliza los valores, conocidos como estados, de estas columnas para predecir los estados de una columna que se designa como elemento de predicción. Específicamente, el algoritmo identifica las columnas de entrada que se correlacionan con la columna de predicción.

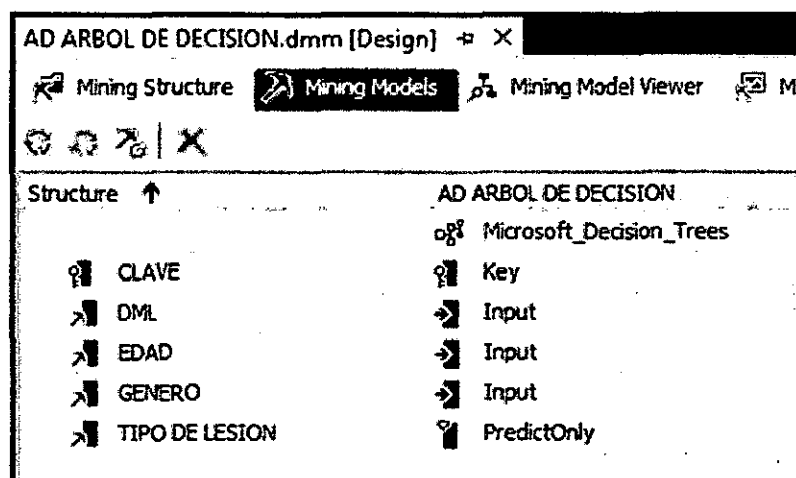
Para los atributos continuos, el algoritmo usa la regresión lineal para determinar dónde se divide un árbol de decisión.

DATOS REQUERIDOS PARA LOS MODELOS DE ARBOLES DE DECISIÓN

Los datos requeridos para los árboles de decisión, incluidos el volumen de datos necesario y la forma en que estos se utilizan son los siguientes:

- **Una columna key:** cada modelo debe contener una columna numérica o de texto que identifique cada registro de manera única. No están permitidas las claves compuestas. Para nuestro modelo se ha creado una columna denominada clave de tipo numérica entera y correlativa.
- **Una columna de predicción:** Se requiere al menos una columna de predicción. Puede incluir varios atributos de predicción en un modelo y pueden ser de tipos diferentes, numérico o discreto. Sin embargo, el incremento del número de atributos de predicción puede aumentar el tiempo de procesamiento. Para nuestro modelo se ha utilizado como columna de predicción el Tipo de Lesiones.
- **Columnas de entrada:** Se requieren columnas de entrada, que pueden ser discretas o continuas. Aumentar el número de atributos de entrada afecta al tiempo de procesamiento. Para nuestro modelo se han utilizado como columnas de entradas el Género, la Ubicación Geográfica y la Edad.

Fig.4.27 - Datos del Modelo Árboles de Decisión



CASO DE INVESTIGACIÓN

Se planifica un escenario para predecir el Tipo de Lesión, de acuerdo a la edad, el género, la ubicación geográfica desde el mes de enero del año 2006 hasta el 30 de abril del año 2015 obteniéndose el modelo de la Fig. 4.27. Los resultados del modelo se analizan con ayuda en la intensidad de los colores para su rápida interpretación, del cual podemos inferir lo siguiente: De los 61975 registros de lesiones, que constituyen el 70% del total de escenarios, 32642 escenarios son de género masculino Fig. 4.28 y 29301 son de género femenino Fig. 4.29 en el primer nivel.

Fig.4.27 - Modelo de Árbol de Decisión

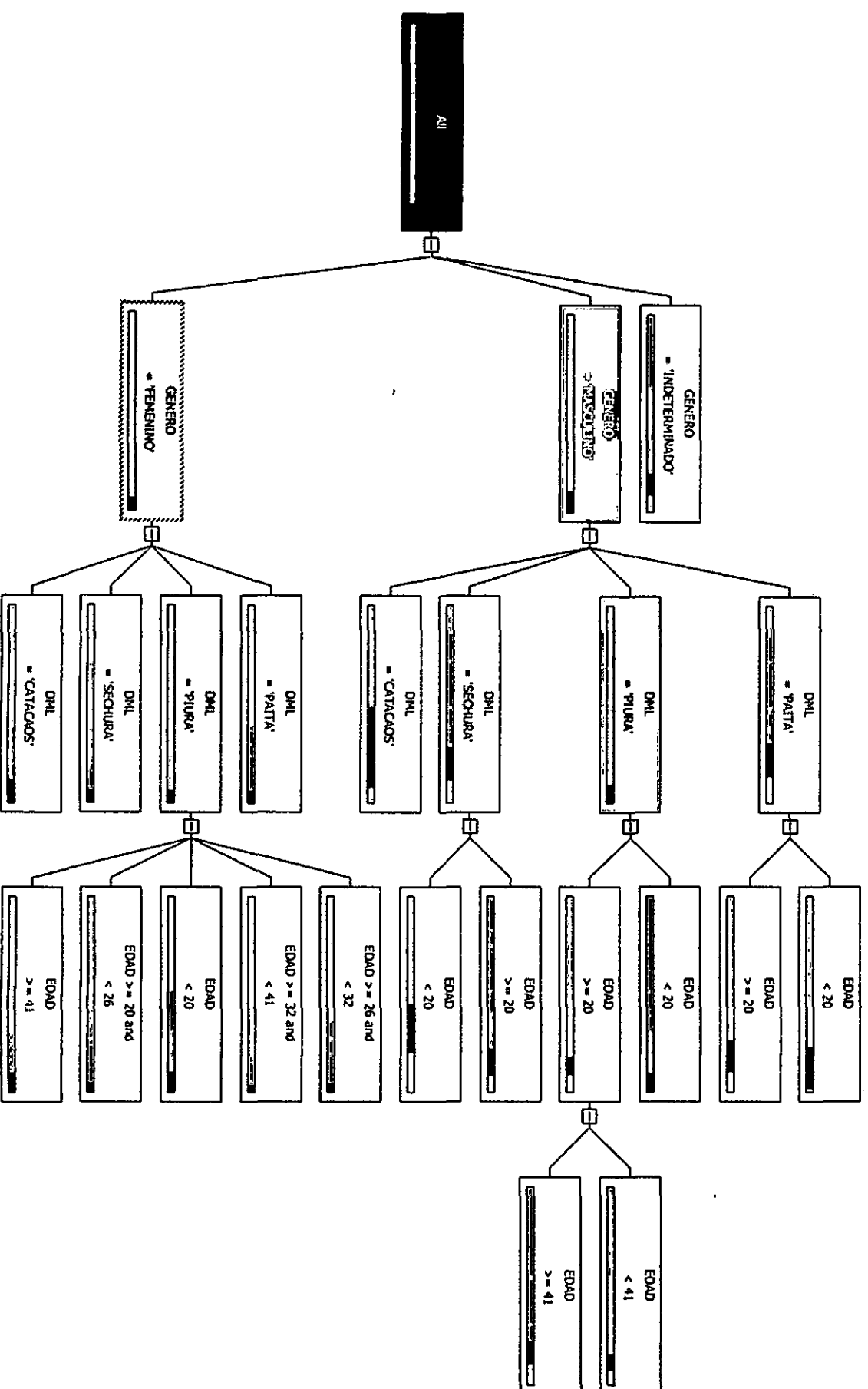
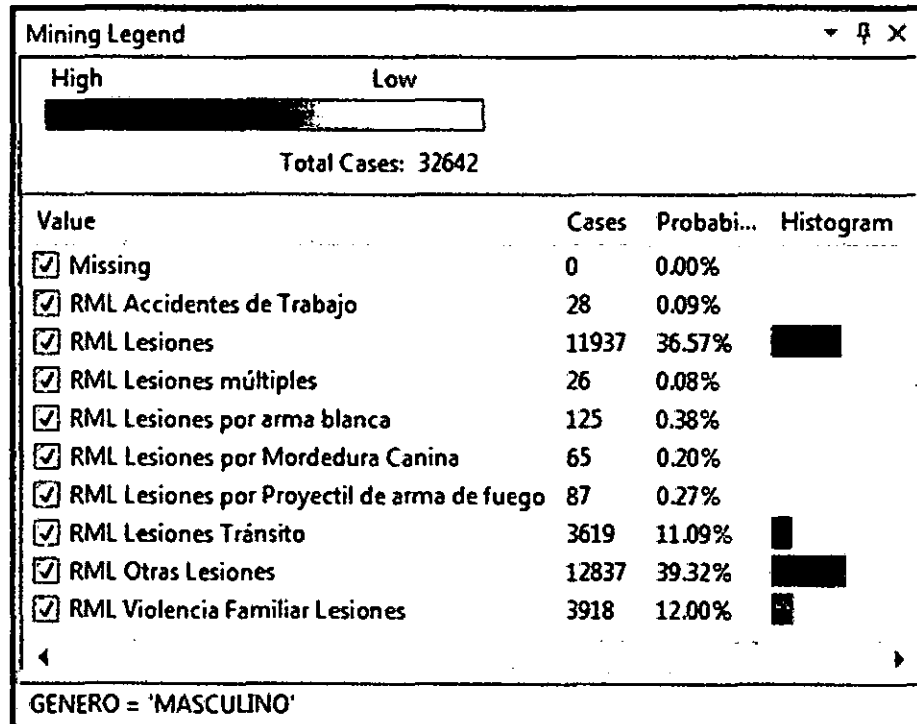


Fig. 4.28 Histograma Tipos de Lesiones (Género Masculino)

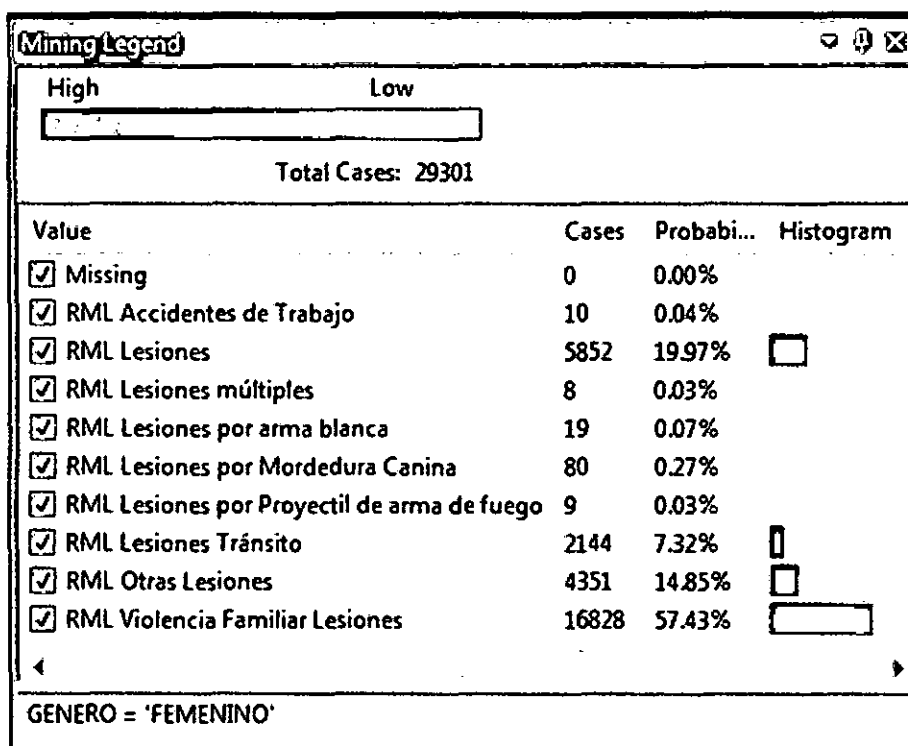


INTERPRETACIÓN

De acuerdo al modelo obtenido se han encontrado patrones de comportamiento los cuales predicen que en el Género Masculino el tipo de lesión más probable que se puede dar en primer lugar es otras lesiones, en segundo lugar lesiones, en tercer lugar violencia familiar lesiones y en cuarto lugar lesiones de tránsito, luego siguen otros tipos de lesiones en menor porcentaje.

En el apartado 4.5.1. la parte donde dice "DESCRIPCIÓN DEL MODELO DE BAYES NAIVE E INTERPRETACIÓN DE RESULTADOS" ítem a) del segundo caso de análisis lesiones por suceso de tránsito, el resultado coincide con la Fig. 4.28 y Fig. 4.29 del modelo de Árboles de Decisión donde el género masculino es el más afectado por casos de lesiones por suceso de tránsito y donde existe una mayor probabilidad de que se registre un nuevo caso por este tipo de lesión.

Fig. 4.29 Histograma Tipos de Lesiones (Género Femenino)



INTERPRETACIÓN

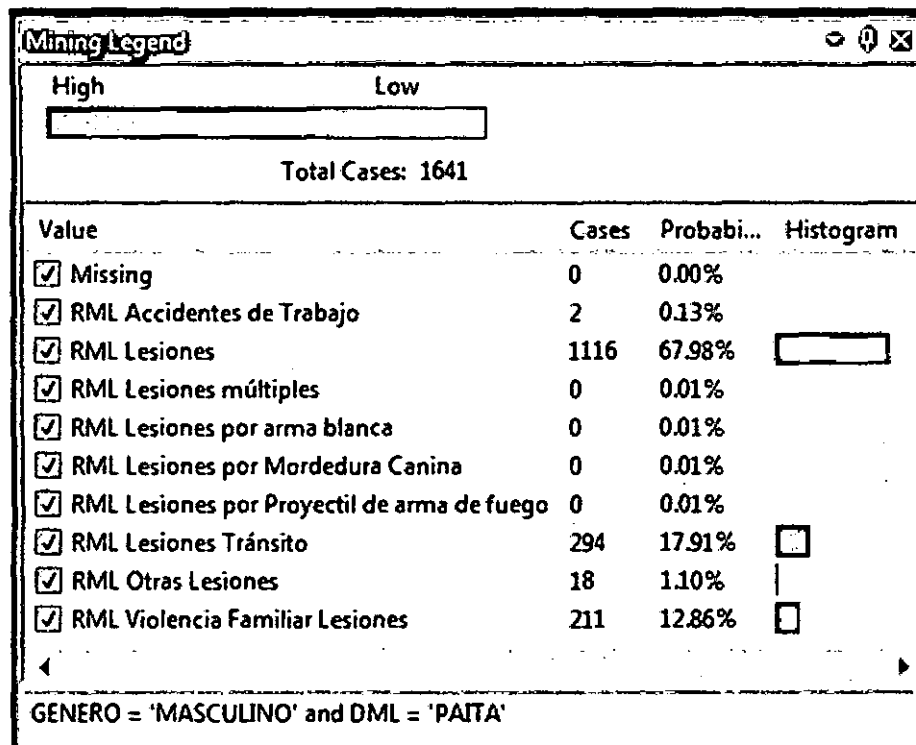
De acuerdo al modelo obtenido se han encontrando patrones de comportamiento los cuales predicen que en el Género Femenino el tipo de lesión más probable que se puede dar en primer lugar es lesiones por violencia familiar, en segundo lugar lesiones, en tercer lugar otras lesiones y en cuarto lugar lesiones de tránsito, luego siguen otros tipos de lesiones en menor porcentaje.

En el apartado 4.5.1. la parte donde dice "**DESCRIPCIÓN DEL MODELO DE BAYES NAIVE E INTERPRETACIÓN DE RESULTADOS**" ítem a), el resultado coincide con la Fig. 4.29 del modelo de Árboles de Decisión donde el género femenino es el más afectado por casos de lesiones por violencia familiar.

En un segundo nivel del árbol de decisión obtenido como modelo encontramos los nodos por ubicación geográfica (DML), teniendo como camino el género Masculino y la zona de Paita (Fig. 4.30) tenemos 1641 casos, con la zona de Piura (Fig. 4.31), tenemos

29068 casos, con la zona de Sechura (Fig. 4.32) 1186 casos y con la zona de Catacaos (Fig. 4.33) 747 casos.

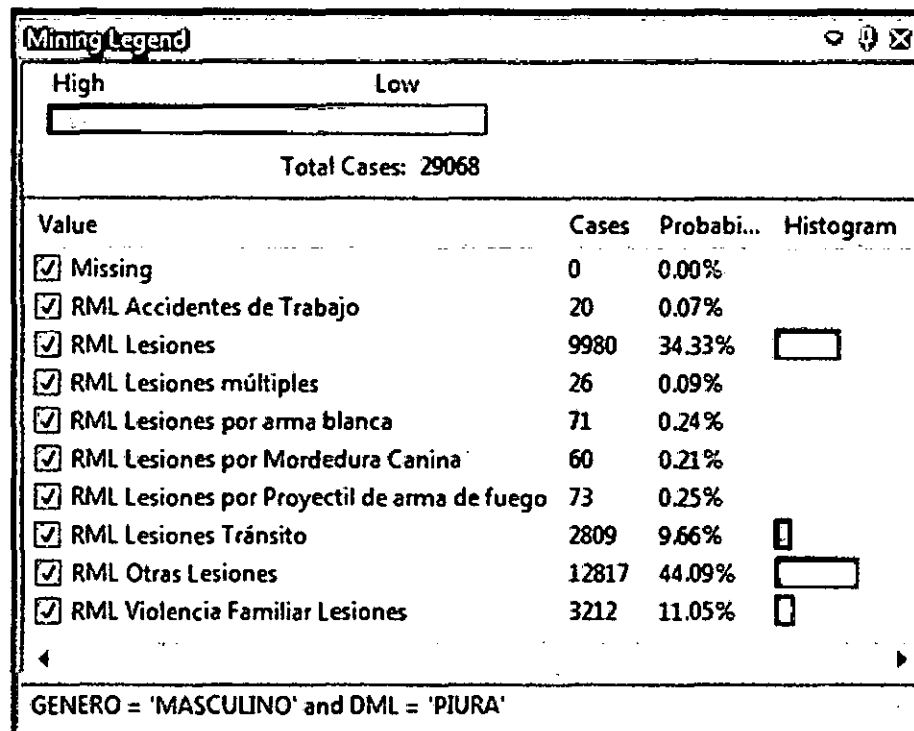
Fig. 4.30 Histograma Tipos de Lesiones
Género Masculino - Zona Paita



INTERPRETACIÓN

De acuerdo al modelo obtenido se han encontrado patrones de comportamiento los cuales predicen que en el Género Masculino y en la zona de Paita, el tipo de lesión más probable que se puede dar en primer lugar es lesiones, en segundo lugar lesiones por suceso de tránsito, en tercer lugar violencia familiar lesiones y en cuarto lugar otras lesiones, luego siguen otros tipos de lesiones en menor porcentaje.

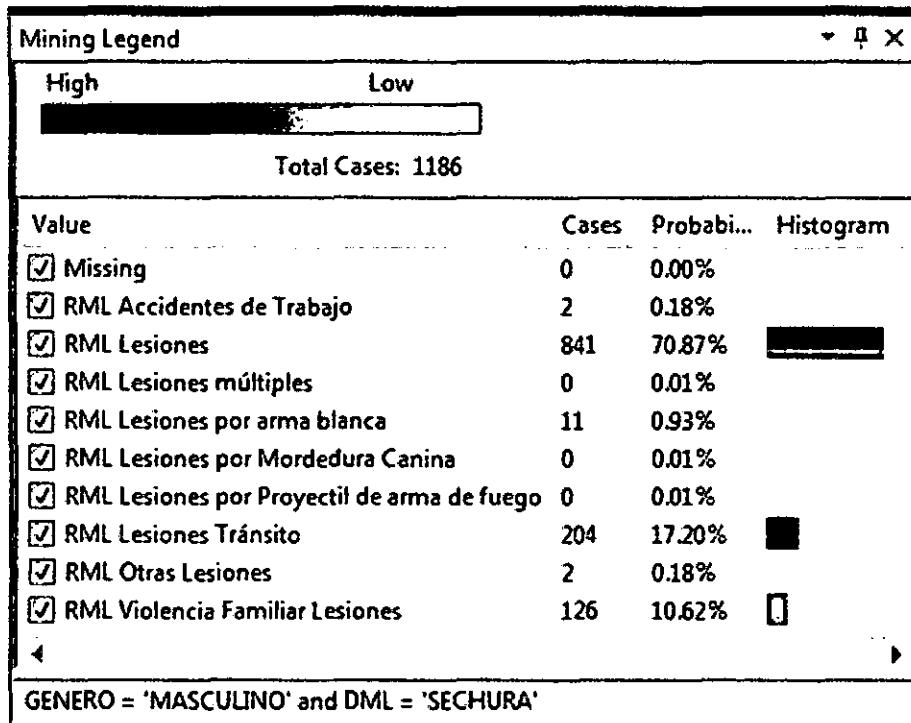
**Fig. 4.31 Histograma Tipos de Lesiones
Género Masculino - Zona Piura**



INTERPRETACIÓN

De acuerdo al modelo obtenido se han encontrado patrones de comportamiento los cuales predicen que en el Género Masculino y en la zona de Piura, el tipo de lesión más probable que se puede dar en primer lugar es otras lesiones, en segundo lugar lesiones, en tercer lugar violencia familiar lesiones y en cuarto lugar lesiones por suceso de tránsito, luego siguen otros tipos de lesiones en menor porcentaje.

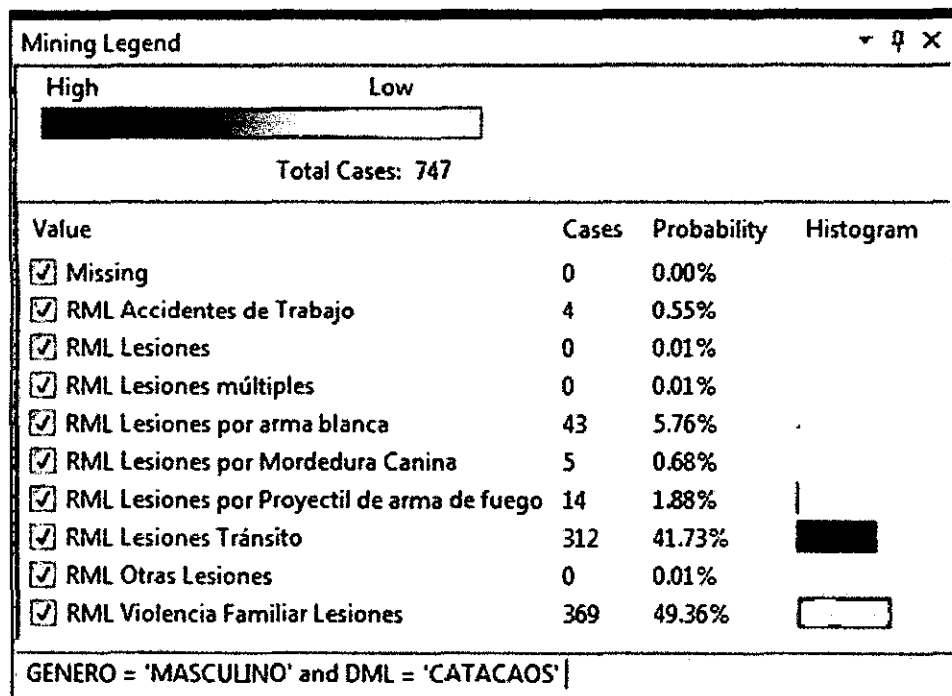
Fig. 4.32 Histograma Tipos de Lesiones
Género Masculino - Zona Sechura



INTERPRETACIÓN

De acuerdo al modelo obtenido se han encontrando patrones de comportamiento los cuales predicen que en el Género Masculino y en la zona de Sechura, el tipo de lesión más probable que se puede dar en primer lugar es lesiones, en segundo lugar lesiones por suceso de tránsito, en tercer lugar lesiones por violencia familiar y luego siguen otros tipos de lesiones en menor porcentaje.

Fig. 4.33 Histograma Tipos de Lesiones
Género Masculino - Zona Catacaos



INTERPRETACIÓN

De acuerdo al modelo obtenido se han encontrado patrones de comportamiento los cuales predicen que en el Género Masculino y en la zona de Catacaos, el tipo de lesión más probable que se puede dar en primer lugar es lesiones por violencia familiar, en segundo lugar lesiones por suceso de tránsito, en tercer lugar lesiones por arma blanca, en cuarto lugar lesiones por proyectil de arma de fuego y luego siguen otros tipos de lesiones en menor porcentaje.

De la Fig. 4.15 y la Fig. 4.33 se puede concluir el mismo resultado obtenido del modelo planteado con el algoritmo Bayes Naive mediante la consulta singleton propuesta en la Fig. 4.13 y el modelo planteado con el algoritmo Árboles de Decisión.

COMPARACIÓN DE MODELOS PARA ANALIZAR EL TIPO DE LESIÓN POR VIOLENCIA FAMILIAR

Fig. 4.34 Histograma Tipos de Lesiones
Modelo Árboles de Decisión

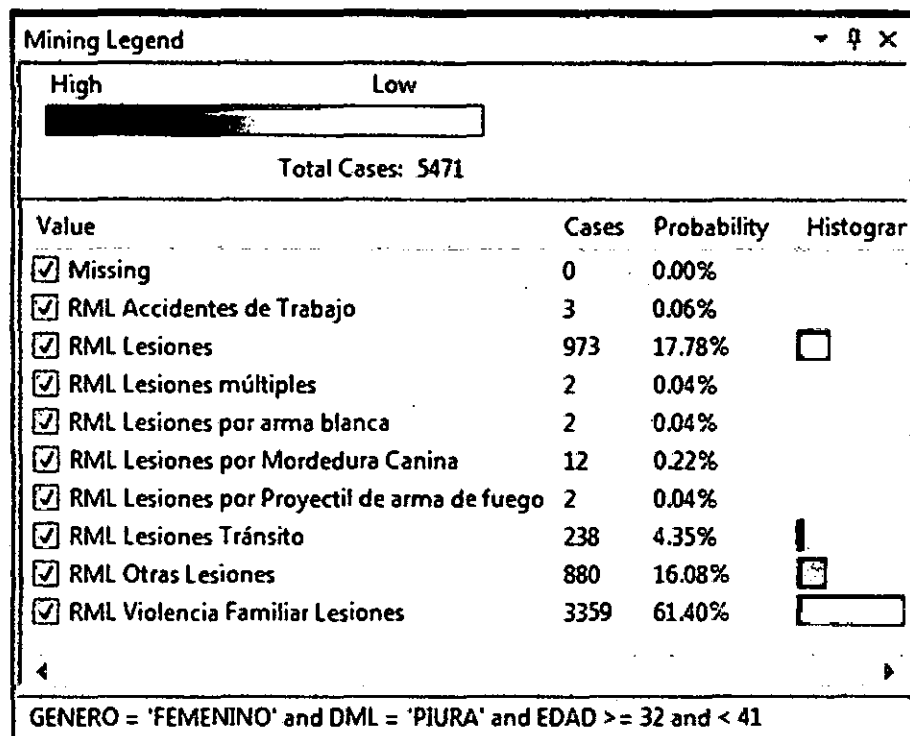
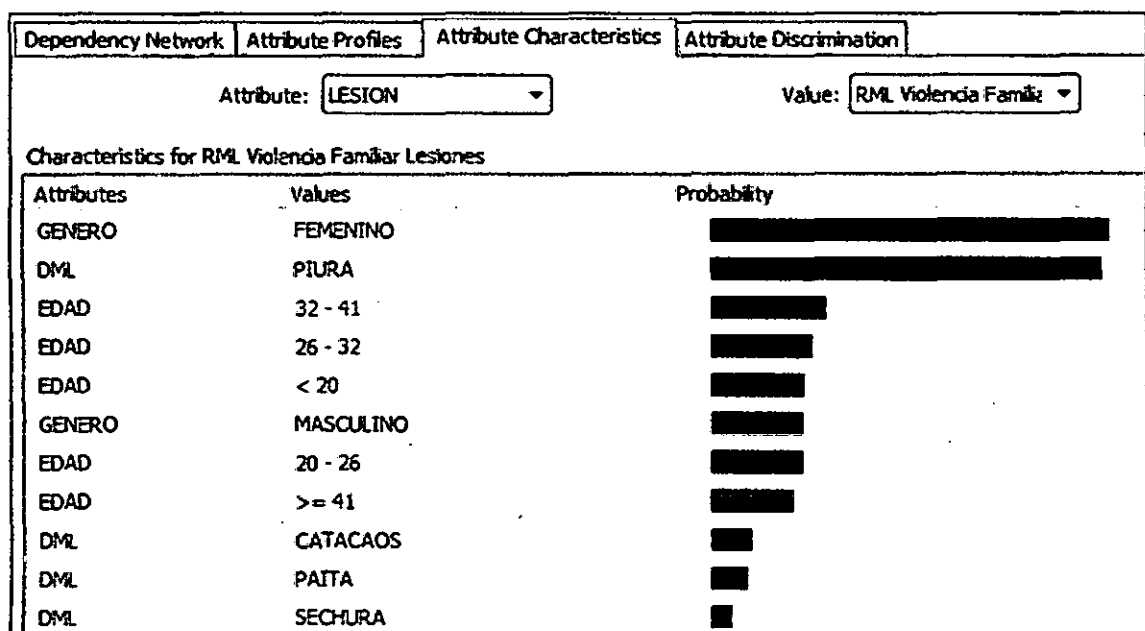


Fig. 4.35 Histograma Lesiones por Violencia Familiar
Modelo Bayes Naive



INTERPRETACIÓN

De acuerdo al modelo obtenido por el algoritmo de árboles de decisión con el histograma que se presenta en la Fig. 4.34 donde se analizan los tipos de lesiones para el género femenino, en la zona de Piura y edad entre 32 a 41 años, y de acuerdo al modelo obtenido por el algoritmo de Bayes Naive con el histograma que se presenta en la Fig. 4.35 donde se analiza el tipo de lesión por violencia familiar, en ambos modelos se observa que la mayor probabilidad que ocurra un caso de lesión por violencia familiar se daría en las mujeres que viven en Piura y con una edad que oscila entre los 32 a 41 años.

CONCLUSIONES

1. Se Implementó el algoritmo de Bayes Naive, el algoritmo de Series de Tiempo y el algoritmo de Árboles de Decisión sobre los datos registrados en la División Médico Legal del Distrito Fiscal de Piura para analizar los diferentes tipos de lesiones.
2. Con los algoritmos de Bayes Naive, Series de Tiempo y Árboles de Decisión se encontraron patrones de comportamiento en los datos registrados de los diferentes tipos de lesiones y se pueden realizar predicciones con los modelos obtenidos.
3. Con el algoritmo de series de tiempo y los modelos obtenidos se pueden pronosticar en el tiempo los diferentes tipos de lesiones, en el presente Proyecto de Tesis se ha pronosticado las lesiones por violencia familiar, las lesiones por sucesos de tránsito, las lesiones producidas por arma blanca y las lesiones producidas por proyectil arma de fuego, con resultados descritos en el capítulo Resultados y Discusión.
4. De acuerdo al modelo de árboles de decisión se han encontrado patrones de comportamiento los cuales predicen que en el Género Femenino el tipo de lesión más probable que se puede dar en primer lugar es lesiones por violencia familiar, resultado que coincide con el modelo de Bayes Naives.
5. De acuerdo al modelo obtenido por el algoritmo de árboles de decisión donde se analizan los tipos de lesiones para el género femenino, en la zona de Piura y edad entre 32 a 41 años, y de acuerdo al modelo obtenido por el algoritmo de Bayes Naive donde se analiza el tipo de lesión por violencia familiar, en ambos modelos se observa que la mayor probabilidad que ocurra un caso de lesión por violencia familiar se daría en las mujeres que viven en Piura y con una edad que oscila entre los 32 a 41 años.

RECOMENDACIONES

1. Se recomienda aplicar el concepto de escalabilidad continuando con el desarrollo de otros algoritmos de minería de datos de Analysis Services o de otros software similar con la finalidad tomar mejores decisiones al verificar que los resultados de los diversos algoritmos utilizados convergen en el mismo resultado.
2. Se recomienda ampliar las fuentes de datos con lesiones registradas en otros Distrito Fiscales del País, de esta forma se busca ampliar el análisis de los tipos de lesiones por parte de las instituciones del Estado Peruano que guardan relación con el tema investigado.
3. Se recomienda desarrollar un plan de capacitación sobre el uso y la interpretación de la información proporcionada por los algoritmos de Minería de Datos sobre todo en su funcionamiento y aplicación, de esta forma los usuarios de la información podrán tomar mejores decisiones. También deben estar incluidos en este Plan de Capacitación el o los desarrolladores de los modelos de Minería de Datos sobre nuevas herramientas que existen o pueden aparecer en el mercado.
4. Las instituciones del Estado Peruano como el Ministerio Público deben permanentemente innovar su plan de tecnologías de la información con nuevas herramientas y técnicas que permitan generar valor en la organización, al mismo tiempo la tecnología debe ser el soporte al plan de negocios de la institución.

BIBLIOGRAFIA

1. BIBLIOTECA msdn (en español). [en línea]: catálogo automatizado de la Biblioteca msdn. Disponible en: <<http://msdn.microsoft.com/library/>>. Contiene: "MSDN Library en español, una fuente de información esencial para desarrolladores que utilizan herramientas, productos, tecnologías y servicios de Microsoft®. MSDN Library incluye procedimientos y documentación de referencia, código de ejemplo, artículos técnicos, entre otros recursos"
2. CHILQUE, Víctor E. (sin fecha). Inteligencia de Negocios. [en línea]. Facultad de Ingeniería y Arquitectura de la Universidad San Martín de Porras. Disponible en: <<http://www.usmp.edu.pe/publicaciones/boletin/fia/info46/sistemas/articulo6.htm>> [2010, Octubre 10]. Contiene: "Artículo de Inteligencia de Negocios del Área de Ebusiness y Proyectos de Tecnología de Información de la Institución arriba mencionada".
3. EGUILA, Arturo y PARCO, Alex A. Implementación de una herramienta de inteligencia de negocios para la administración de justicia sobre una metodología ad-hoc. Tesis (Ingeniero de Sistemas e Informática). Lima, Perú, Universidad Nacional Mayor de San Marcos, Facultad de Ingeniería de Sistemas e Informática, 2007. 120 p. Disponible en: <http://www.cybertesis.edu.pe/sdx/sisbib/fiche.xsp?base=documents&id=sisbib.2007.parco_ja-principal>. Contiene: "Metadatos del documento de Tesis, y, enlace al documento mismo arriba descrito".
4. ESPINOZA, Dhony C. y QUISPE, Renan R. Solución de inteligencia de negocios para empresas de servicios de asistencia aplicación práctica a la gerencia de asistencia del Touring y Automóvil Club del Perú. Tesis (Ingeniero de Sistemas e Informática). Lima, Perú, Universidad Nacional Mayor de San Marcos, Facultad de Ingeniería de Sistemas e Informática, 2006. 112 p. Disponible en: <http://www.cybertesis.edu.pe/sisbib/2006/quispe_ar/pdf>. Contiene:

"Listado de directorio de los documentos de Tesis de los autores antes mencionados en este ítem".

5. GOPACSI (sin fecha). MundiHogar. En Historias de éxito [en línea]. Disponible en: < <http://www.gopac.com.mx/v3/empresa/historias/mundihogar.asp> > [2010, Setiembre 12]. Contiene: "Sitio web de una empresa desarrolladora de software especializada en desarrollar soluciones de inteligencia de negocios para otras empresas".
6. GOPACSI (sin fecha).. CPIngredientes obtiene el Poder de la Información. En Historias de éxito [en línea]. Disponible en: < <http://www.gopac.com.mx/tribune/Historias/CPI/cpi.htm> > [2010, Setiembre 12].
7. SINNEXUS. (sin fecha). Bases de datos OLTP y OLAP. En ¿Qué es Business Intelligence? [en línea]. Disponible en: <http://www.sinnexus.com/business_intelligence/olap_vs_oltp.aspx> [2010, Octubre 6]. Contiene: "Sitio web de empresa Sinnexus de nuevas tecnologías, cuya principal fuerza impulsora es la aportación de productos y servicios de inteligencia de negocio (business intelligence) y sistemas de información geográfica (GIS)".
8. SUPERINTENDENCIA DE BANCA, SEGUROS Y AFP. (sin fecha). Boletín Estadístico de Cajas Municipales. En Sistema Financiero [en línea]. Disponible en: < <http://www.sbs.gob.pe/app/stats/EstadisticaBoletinEstadistico.asp?p=3#>> [2010, Agosto 10]. Contiene: "Sitio web oficial de la Superintendencia de Banca y Seguro del Perú, específicamente el enlace de Boletín Estadístico de Cajas Municipales"
9. WOLFF, Carmen G. (1999). Modelamiento Multidimensional. [en línea]. *Revista Ingeniería Informática*. Concepción, Chile, Universidad de Concepción. Disponible en: < <http://www.inf.udec.cl/~revista/>> [2010,

Setiembre 11]. Contiene: "Enlace de artículos publicados en revista electrónica de Ingeniería Informática".